Predicting Metritis Events in Dairy Cattle Using Machine Learning Classifiers on Multiple Data Streams Under Nowcasting and Forecasting Frameworks.

By

GEMMA VIDAL FABUEL

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Epidemiology

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Beatriz Martinez-Lopez, Chair

James Sharpnack

Pablo Pinedo

Committee in Charge

2021

Abstract

In recent years an increasing number of precision dairy farming technologies (PDFTs) have been incorporated into the management of dairy operations. Recently, research has been centered on the use of sensors to quantify animal behaviors such as activity level, rumination time, or lying time, and their potential for disease detection. In dairy cows, the transition period around parturition is considered the time when most diseases occur, being hypocalcemia, metritis, or hyperketonemia the most common. Whether the combination of different behaviors registered by sensors can better diagnose diseases during transition period, their sensitivity and specificity to detect diseases, or what is their predictive ability or how far in advance they can detect disease is not known. Our goal was to develop, test and validate a workflow for disease surveillance in dairy cattle with emphasis on metritis, using a combination of feature selection strategies and machine learning algorithms. The long-term goal was to provide a framework where high-frequency time series behavioral data registered by multiple PDFTs could be used as a tool for early detection of dairy cattle health problems during the transition period. Data from 35 dairy cows that either did not experience any disease postpartum or were only diagnosed with metritis were retrospectively selected from a dataset containing behavioral, production, and clinical data from 138 lactating cows during the first 21 days postpartum at the University of Kentucky Coldstream Dairy (Lexington, KY, USA). Metritis events were created based on changes in metritis scores recorded during clinical examination. After a review of PDFTs and machine learning approaches (Chapter 1), Chapters 2 and 3 study the classification performance of three classifiers (*k*-nearest neighbors, random forest, and support vector machines) when predicting

metritis events by using behaviors registered by two different 3-axis accelerometers. Chapter 4 studies the classification performance of a random forest classifier to predict metritis events when multiple inputs from multiple data streams were combined. Multiple time windows, time lags, and classification thresholds were compared under nowcasting (Chapter 2, 3, and 4) and forecasting frameworks (Chapter 4). Random Forest had the greatest F_1 score across time windows and time lags, but best behaviors for classification changed depending on the combination of time window and time lag. Furthermore, forecasting metritis events 2 and 3 days forward had similar performance results compared to the nowcasting framework. Based on our findings, machine learning classifiers can aid in the identification of animals at higher risk of being sick before traditional diagnosis is performed.

Acknowledgments

Many times during my PhD I visualized myself going through this very moment, thanking all the people I have to thank for their support. There is always a difference between what we have been imagining a moment is going to be like, and reality. Not even in my wildest dreams I could have imagined I would be graduating during a global pandemic, with a move across the Atlantic Ocean in the mix. The proverb says it takes a village to raise a child, but it turns out it also takes a village to get a PhD. So here is my village.

I would like to express my sincere gratitude to my dissertation committee Beatriz Martínez-López, James Sharpnack, and Pablo Pinedo for walking this path along with me. As my major professor, Beatriz has always supported me in my exploration process and development as researcher. She supported me in my decision of exploring classes from other departments, something that brought me the opportunity to collaborate with other professors such as James. Who would have thought when I walked into James' office for the first time, asking him if I should drop his class, that I was going to leave his office with a new dissertation committee member (and an extra class for that quarter)? I am thankful because during these years I've found in James' office a safe space to chat, laugh, vent, or think about how to translate my thoughts into code with my gaze lost in the office ceiling. Pablo has also been an invaluable support for me during this time. I met Pablo in 2009, when I was an intern while he was a resident at the University of Florida. We shared many bonding experiences during our clinical training program, and we have been in touch ever since. I am glad Pablo accepted being in my dissertation committee, bringing his expertise in dairy cattle. Also, I would like to thank my

qualifying exam committee members, Heejung Bang, Joanne Rowe, Beate Crossley, James Sharpnack, and Woutrina Smith for providing constructive guidance that was key during the initial phase of my dissertation. I would also like to thank Jeffrey Bewley for facilitating all the data sharing between his former students and me. Students at University of Kentucky and Coldstream Dairy staff did a great job at collecting all the data during multiple years and I ended up using their datasets in all the dissertation chapters.

At the time I was starting my PhD I realized that the finances of doing a PhD were going to be a constant struggle, and the times I thought about giving up on my PhD had to do with financial vulnerability. I am grateful to everybody that contributed by providing funds, directly or indirectly, towards my PhD. Eric and Cindy Davis from R-Vets have been invaluable to me. Besides their financial support, they provided me with opportunities for personal and professional growth through a series of trips to Eastern Nicaragua. During moments of acute financial stress, Wendi Jackson always vouched for me, helping me to secure a GSR position working for the CAS-Pakistan project, an opportunity that allowed me to work with Pat Conrad, Alan Conley, Jim Hill and the International Projects Office. I am grateful to all of them for taking me in and mentoring me. It was my first experience working with an international project management team and I learnt a lot from them. Upon the end of the CAS-Pakistan project, I started (again thanks to the mediation of Wendi Jackson) working with Woutrina Smith for the YSM and PREDIC-2 projects. Woutrina supported me throughout the years I worked for these projects, and even though she was not my major professor or advisor, Woutrina took care of me the way a major professor would take care of her students, not only financially but also by providing me with learning opportunities and building up my tool kit as a graduate student. I also

want to thank Fernanda Ferreira for meeting with me and listening to my research ideas. Thanks to her leadership and expertise, funding came our way during the last year of my PhD. The YSM and PREDICT-2 projects were part of the One Health Institute at UC Davis, where I had the opportunity to work with David, Ian, and Jenny. Even though working on a project in Ethiopia had its challenges, and traveling back and forth was tiring, traveling with Jennie made those trips memorable and I would do it all over again in a heartbeat. Other sources of funding came through the Graduate Group in Epidemiology and as teaching assistantships from the Animal Science Department and the School of Veterinary Medicine.

During the process of getting my PhD there have been many ups and downs. I would like to thank to the people that put the energy into listening to me and to walk on my shoes: Phil Kass, Lorena García, Tami Ali from the GGE, and Daniel Moglen from the FUTURE program.

I've been a veterinarian for a long time now and I consider myself very lucky for all the people that has had a mentoring role throughout my career. They are these sort of unintentional life coaches that saw my potential where I couldn't see much and gave me that small but necessary push when I needed to be pushed in order to take on new career challenges. I've been able to get out of my comfort zone and keep growing professionally thanks to Cristòfol Peris and his guidance during my first steps as researcher, setting the bar of what a mentor should be really high, Ángel García for being there for me during my struggles navigating adulthood and giving me the last push to move to the U.S., Pepe García, Andrés García, and Antonio Casas for their guidance during my first steps as a livestock clinician, Jose Santos for taking me in when I first came to the U.S. and for pushing me to pursue a residency program in the U.S., Carlos Risco, Owen Rae, Art Donovan, Joanne Rowe, Bruce Hoar, and Mike Lane, for believing in me as

clinician and researcher while seeing me as a whole person. And to Nacho, of course, for making me believe that the world was my oyster, even though our marriage did not survive the process.

My life wouldn't be the same without my friends. They have been an incredible emotional support system for me and we have supported each other the way a family would. I am grateful for having in my life people like Amanda and her unconditional support during my life struggles, Fernando, Shalini, Kwabena, Leah, Edward, Anne-Marie, Val, Ale, Melissa, Jennie, Nistara, Pranav, Megan, Lucho, Karla, Vona, Mulu, Lore, Alejo, Karinna, Marco, Heather, Felipe, and Lupita, my friends from the Graduate Group in Epidemiology: Laura, Wendi, Monica, Lauren, Tanya, Gaby, and to my friends from CADMS lab: Jaber, Kyu, Jerome, Becky, Inés, and Nacho.

And last but not least, I am thankful to my family: my parents, my sister, Tada, and Martín, for their unconditional love and support even though I keep pushing them out of their comfort zone.

Table of Contents

List of Tables

List of Figures

Frequently Used Abbreviations

5-FCV	Fivefold cross-validation
AMS	Automated milking system
AUC	Area under the curve
BHBA	Beta-hydroxybutyrate
DIM	Days in milk
DT	Decision trees
FN	False negatives
FP	False positives
GS	Grid search
k-NN	*k*-nearest neighbors
ML	Machine learning
MSE	Mean squared error
NEFA	Non-esterified fatty acid
NPV	Negative predictive value
PDFT	Precision dairy farming technology
PPV	Positive predictive value
PR-curve	Precision-Recall curve
RF	Random forest
ROC	Receiver operating characteristics
RS	Randomized search
SCC	Somatic cell count
Se	Sensitivity
Sp	Specificity
SVM	Support vector machines
TN	True negatives
TP	True positives

1 Introduction

1.1 Background Information

Syndromic surveillance can be defined as the real-time (or near real-time) collection, analysis, interpretation, and dissemination of health-related data to enable the early identification of the impact (or absence of impact) of potential human or veterinary public health threats which require effective public health action (Dupuy et al., 2013). Modern biosurveillance systems are designed to take advantage of data related with health conditions that, once grouped into syndromes, can allow for faster detection of health problems despite being less specific than traditional diagnostic methods (Dórea et al., 2013). Previous studies have used different sources of data such as clinical records, laboratory submissions or production data in order to explore the potential different types of data have for syndromic surveillance. From all those, data recorded individually for each animal regarding different aspects of productivity and well-being offers the shortest time lag between a health event and its potential detection (Dórea et al., 2016).

For the past years, dairy farms have been increasing in size, resulting in a lower ratio of labor available per animal making difficult systematic checking of animals at risk of disease on a regular basis (Paudyal et al., 2016). At the same time, there has been an increasing number of precision dairy farming technologies (PDFTs) used by the dairy industry in order to increase overall farm production efficiency (Wathes et al., 2008). These new automated data collection

systems generate large amounts of information at a higher frequency that could be used for syndromic surveillance of a variety of syndromes in dairy cows.

Sickness behavior is part of an adaptive response to infection or injury that occurs when the animal is trying to cope with a stressor. Most of the sickness behavior are associated with depression, loss of appetite, weight loss, and pain (Tizard, 2008). However, traditional observation of sickness behavior on farms are often based upon subjective clinical evaluation, which are in turn influenced by the accumulated experience of farm personnel, with questionable consistency (Espadamala et al., 2016). The availability of new technologies to automatically record behaviors allows for increased use of objective measurement, and changes in behavioral patterns can be used to predict, identify, and assess health problems at the individual animal-level to prevent or mitigate clinical disease (Weary et al., 2009; LeBlanc, 2010; Dittrich et al., 2019). Recent PDFT literature has been focused on the validation of these technologies regarding feeding, rumination, and lying behaviors (Bikker et al., 2014; Borchers et al., 2016). However, despite the claims of manufacturer companies about other applications such as diagnosis of ketosis, mastitis, reproductive status or protein status, the accuracy and prediction ability of these PDMTs for many livestock disease syndromes is still uncertain.

Due to metabolic challenges, diseases such as hypocalcemia, hyperketonemia, and metritis are commonly diagnosed around the transition period in 30 to 50% of dairy cattle (LeBlanc, 2010). These health events have effects on welfare and on short and long term reproductive health (LeBlanc, 2010). Despite the impact of diseases during the transition period, in a review by Rutten et al. (2013), only 16% of the precision dairy farming literature were related with disease around parturition. Furthermore, common limitations of these studies were

the lack of control for concurrent postpartum diseases, behavioral data aggregations before and after disease diagnosis resulting in lost temporal relationships, and lack of consideration of within-same-day behavior variability due to farm scheduled activities (Huzzey et al., 2007a; Stoye et al., 2012).

1.2 Precision Dairy Farming Technologies

Precision dairy farming involves the use of technologies or biosensors to measure physiological, behavioral, and production indicators on individual animals. The primary goals of precision dairy farming are to 1) maximize animal performance, 2) detect diseases in individual cows early, 3) detect herd level health and production problems early, and 4) minimize the use of medication through preventive health measures (Bewley, 2010). The use of PDFTs enables the fulfillment of these goals without too much additional labor input (Bewley, 2010). By processing the data collected by PDFTs in combination with decision support systems, the application of these sensors improves animal monitoring of a wide range of conditions such as heat stress, postpartum diseases, lameness, and mastitis (Lee, 2018).

Technologies fall into two categories: the device is on or inside the cow's body (cow-attached PDFTs), or the device is off the cow's body and measurements are registered as the cow walks past or through the device, or a sample is taken to run an analysis (non-attached PDFTs; Tsai, 2017). Developed sensor systems can be divided into four different levels: (I) sensor technique or signal processing of the changes in the sensor and their assumed relation with the animal's behavior (e.g. steps); (II) integration of information and measurement of changes to

generate information about the cow's status (e.g. increased activity and estrus); (III) integration of sensor information with other sources of information to generate advice at the individual (e.g. whether to inseminate a cow or not) or herd level if data are aggregated; (IV) decision making autonomously by the sensor system (Rutten et al., 2013). Most sensor systems are at levels I, II, and III, where companies develop proprietary algorithms to generate alerts when an increased likelihood of a disease event is detected for a given animal (Tsai, 2017).

Different levels require different validation methods. At level I, the classification of the animal behavior by the sensor can be achieved through different statistical methods (e.g., machine learning classifiers), and the results are usually validated against visual observations. At level II, through the analyses of one or multiple behaviors measured by one or multiple PDFTs, associations with diseases or reproductive events are estimated using a variety of statistical methods (e.g., machine learning classifiers). Validation is made through traditional clinical diagnosis of disease or reproductive events such as visual observations followed by additional tests, using blood, urine, or milk samples (Sepúlveda-Varas et al., 2014).

During validation of a PDFT, different performance metrics are used. These are of great importance in order to be able to judge the usefulness of a given PDFT and to compare this to competing PDFTs. Basic measurements included for binary outcomes (sick and non-sick) are the frequency of true positives (TP; number of observations that are positive and identified by the test as positive), false positives (FP; number of negatives and identified as positives by the test), true negatives (TN; number of observations that are negative and identified by the test as negative), and false negatives (FN; number of positives and identified as negatives by the test). Based on these, other metrics such as sensitivity (Se), specificity (Sp), positive predictive value

(PPV), negative predictive value (NPV), and accuracy (Ac) can then be calculated (Iwersen et al., 2009; Hogeveen et al., 2010). Briefly, Se is calculated as TP / (TP + FN), Sp as TN / (TN + FP), PPV as TP / (TP + FP), NPV as TN / (FN + TN), and Ac as (TP + TN) / (TP + FP + TN + FN). These are considered single-threshold measures, because they are defined for individual score thresholds or cutoff of a test, and cannot give an overview of the range of performance with varying thresholds (Dominiak and Kristensen, 2017). A solution to this problem is to use the Area under the Receiver Operating Characteristics (ROC) curve (AUC), where it shows pairs of Se and Se values calculated at all possible thresholds (Saito and Rehmsmeier, 2015). A summary of performance in validation studies for cow-attached and non-attached PDFTs with emphasis on diseases postpartum (hyperketonemia, hypocalcemia, metritis, or the combination of the three) can be found in Tables 1.1 and 1.2. However, the reader should be aware that due to the large variation in reported validation methods, test scales, and algorithms used, it is difficult to compare the performance of various sensor types across studies.

Sickness behavior is part of an adaptive response to infection or injury that occurs when the animal is trying to cope with a stressor (Weary et al., 2009). Most sickness behaviors are associated with depression, loss of appetite, weight loss, and pain (Tizard, 2008). Therefore, these behavioral changes can potentially be used to estimate the risk for diseases (Weary et al., 2009). Changes in behavior could be positive (e.g., increased frequency when cows are sick) or negative (e.g., decreased frequency) when cows are ill (Weary et al., 2009). Traditional observation of sickness behavior on farms are often based upon subjective clinical evaluation, which are in turn influenced by the accumulated experience of farm personnel, with questionable consistency (Espadamala et al., 2016). Common behavioral, physiological, and

production variables measured at the individual animal level by PDFTs are rumination time, feeding, standing, lying time, activity levels, body condition score, heart rate, body temperature, milk yield, and milk components. The availability of these automatically recorded behaviors at high frequency allows for increased use of objective measures, and changes in behavioral patterns can be used to predict, identify, and assess health problems at the individual animal-level to prevent or mitigate clinical disease (Weary et al., 2009; LeBlanc, 2010; Dittrich et al., 2019).

1.2.1 Rumination and Feeding Behavior

Rumination is a natural behavior used to break down the feed particle size and to create a greater concentration of bacteria for fermentation (Russell and Rychlik, 2001). Therefore, it plays a vital role in maintaining high levels of feed intake and efficient digestive function as chewing also helps to increase saliva secretion necessary for an efficient digestive function (Soriani et al., 2012; Beauchemin, 2018). The time a cow spends chewing during either eating or ruminating varies depending on chemical and physical characteristics of the diet, feeding management, and cow variability, but it is also associated with health in dairy cows (Radostits et al., 2006; Beauchemin, 2018). The amount of time cows spend ruminating and eating is highly variable, and it ranges from 151 – 632 min/d and from 141 – 507 min/d, for ruminating and eating, respectively (Beauchemin, 2018). Most rumination occurs at night when cows are at rest, but cattle also ruminate throughout the day when not interrupted by farm feeding schedules or milking hours (Paudyal et al., 2016). During the transition period, feed intake is crucial to manage negative energy balance and to prevent metabolic and infectious diseases postpartum (Urton et

al., 2005). As a consequence, rumination time and feeding behavior are both important variables for the detection of illnesses (Hansen et al., 2003).

Rumination and eating time are being recorded by PDFTs such as CowManager SensoOr (Agis Automatisering, Harmelen, Netherlands), HR Tag (SCR Engineers Ltd., Netanya, Israel), SmartBow (Smartbow GmbH, Jutogasse, Austria), or TrackaCow (ENGS System Innovative Dairy Solutions, Israel). Different devices use various methodologies: CowManager SensoOr is an ear-attached 3-axis accelerometer that records changes associated with jaw and ear movement related to chewing and ruminating (Matsui and Okubo, 1991; Dado and Allen, 1993; Beauchemin, 2018). TrackaCow is also a 3-axis accelerometer but in this case, the device is attached to the cow's leg. Another device is located by the feedbunk and when the device attached to the cow's leg comes in close contact with the device located by the feedbunk, feeding time and number of visits to the feedbunk are registered for a given cow. In contrast, HR Tag is a sensor mounted on a collar with a microphone that captures eructation and rumination sounds. References for validation studies and their findings can be found in Table 1.1.

Changes in rumination and feeding behaviors have been associated with clinical mastitis (Stangaferro et al., 2016b), hyperketonemia, displaced abomasum (Stangaferro et al., 2016a), and metritis (Liboreiro et al., 2015; Stangaferro et al., 2016c; Neave et al., 2018). Cows with metritis were found to ruminate less during a period of time that ranged between 2 to 9 days postpartum (Liboreiro et al., 2015). Cows with clinical metritis ruminated about 50 minutes less per day than healthy cow between 5 days before and 5 days after clinical diagnosis (Stangaferro et al., 2016a), and ate 1 kg/d less during the 3 days before diagnosis (Neave et al., 2018). However, Neave et al. (2018) reported no difference in prepartum feeding time between cows

diagnosed with or without metritis postpartum. In validation studies, ruminating behavior and its association with disease postpartum showed a Se that ranged between 42% to 71%, a Sp between 74% and 96%, and Ac between 73% and 84%, depending on the PDFT and the disease studied (hyperketonemia, hypocalcemia, metritis, or the combination of the three; Table 1.1). Similarly, and also depending on the PDFT used and the disease studied, feeding behavior and its association with disease postpartum showed a Se than ranged between 56% and 79%, Sp between 74% and 91%, and Ac from 74% to 81% (Table 1.1).

1.2.2 Activity and Lying Behavior

In dairy cattle, increased physical activity is a sign of estrus (Firk et al., 2002). Measured with accelerometers that transform acceleration into angles, when attached to the leg, changes in angles are interpreted either as steps or lying. Among behaviors considered as resting states, lying time has a critical role in the production potential and welfare status of dairy cattle, as cows normally need to lie down an average of 12 - 13 hours per day (Drissler et al., 2005; Fregonesi et al., 2007; Gomez and Cook, 2010). In response of a disease, ill cows increase resting time to conserve energy for fever response and activation of the immune system instead of eating or engaging in normal activities (Hart, 1988). Lying behavior could vary by stage of lactation, as lying time and lying bouts increase as days in milk increases (Munksgaard et al., 2005; Vasseur et al., 2012; Ito et al., 2014).

Activity levels and lying behavior are being recorded by PDFTs such as AfiAct Pedometer (Afimilk, Kibbutz Afikim, Israel), IceQube (IceRobotics Ltd., Edinburgh, Schotland), CowManager SensoOr (Agis Automatisering, Harmelen, Netherlands), or TrackaCow (ENGS System Innovative

Dairy Solutions, Israel). Most of these are 3-axis accelerometer attached to the cow's leg or to the cow's ear. Some of these devices report the number of minutes per hour a cow spent performing a behavior that was classified as either not active, active, high activity, while others count the number of steps per hour. In contrast, other devices such as IceQube generate an activity index that is calculated based on a proprietary algorithm. References to validation studies for these devices and their findings can be found in Table 1.1.

Changes in lying time have been used for the detection of diseases during the transition period such as lameness (Proudfoot et al., 2010; Calderon and Cook, 2011), dystocia (Proudfoot et al., 2009), and subclinical hypocalcemia (Jawor et al., 2012). Lame cows have been found to experience longer lying time and lying bouts (Chapinal et al., 2009; Ito et al., 2010). In validation studies, different activity levels and their association with disease postpartum showed a Se that ranged between 53% to 79%, a Sp between 68% and 91%, and Ac between 67% and 81%, depending on the PDFT and disease studied (hyperketonemia, hypocalcemia, metritis, or the combination of the three), with lower Se for those PDFTs that generate their own activity index (Table 1.1). Similarly, and also depending on the PDFT used and disease studied, lying behavior and its association with disease postpartum showed a Se than ranged between 53% and 79%, Sp between 68% and 90%, and Ac from 67% to 81% (Table 1.1).

1.3 Metritis

1.3.1 The Transition Period

In dairy cattle, the periparturient period is characterized by a sudden increase in energy requirements and a decrease in voluntary dry matter intake, creating a temporal negative energy

balance between a faster increment in the energy demand compared with a slower increment in the energy intake (Drackley et al., 2001; Goff, 2006). Increased energy demand is driven by increased metabolic demands by the fetus three to four weeks before calving, and the prioritization of nutrients toward the mammary gland to start a new lactation, causing fatty acid mobilization and the release of non-esterified fatty acids (NEFAs) in order to meet energy requirements (Ingvartsen and Andersen, 2000; Drackley et al., 2001; Overton and Waldron, 2004). The liver, if overwhelmed by an excessive release of NEFAs, will transform these into ketone bodies, exacerbating the negative energy balance due to the clinical signs caused by the hyperketonemia (Baird, 1982). The periparturient period, also called transition period, is defined as the three weeks before and three weeks after parturition (Drackley, 1999). During this period, there is a complex relationship between negative energy balance, increased NEFAs, and immunosuppression, exacerbated by low levels of calcium in blood (Goff, 2006). In fact, most metabolic diseases such as milk fever (hypocalcemia) and ketosis (hyperketonemia) occur within the first 2 weeks postpartum in dairy cows. Even some of the diseases such as laminitis that occur later during the lactation can be traced back to disorders that occurred in early lactation (Donovan et al., 2004). Furthermore, due to the immunosuppression, cows are prone to increased pathogen load (Hammon et al., 2006), with the overwhelming majority of infectious diseases such as metritis, mastitis or salmonellosis becoming clinically apparent during the first 2 week postpartum. It has been estimated that 30 to 50% of dairy cattle are diagnosed with either metabolic or infectious diseases during the transition period (LeBlanc, 2010), and these diseases cause decreased milk production (Fourichon et al., 1999; Edwards and Tozer, 2004; Huzzey et al., 2007a), poor reproductive performance (Opsomer et al., 2000; Walsh et al., 2007), and

increased culling rate (Hadley et al., 2006; Dubuc et al., 2011). Recent studies have estimated that the cost of a case of metritis ranges from $240 to $884 (median $398), and milk price, treatment cost, replacement cost, and feed cost explain 59%, 19%, 12%, and 7%, respectively (Pérez-Báez et al., 2021).

1.3.2 The Uterus During Postpartum

Before parturition the uterine lumen is sterile, but during parturition the cervix opens, allowing environmental contamination to migrate from the vagina to the uterus (Földi et al., 2006). However, bacterial presence does not necessarily assume later uterine infection, as more than 90% of cows have some bacteria present during postpartum that is not associated with clinical disease (Sheldon and Dobson, 2004; Sheldon et al., 2006). Compared with bacterial contamination, infection implies adherence of pathogenic organisms to the mucosa, colonization or penetration of the epithelium, and/or release of bacterial toxins that lead to establishment of uterine disease, which depends on the immune response of the cow and the endocrine environment (Janeway Jr et al., 2001; Sheldon et al., 2006).

During parturition, the expulsion of the fetus occurs along with the associated membranes and fluids. Uterine involution starts right after parturition, involving physical shrinkage, necrosis and sloughing of caruncles, and the regeneration of the endometrium. Expulsion of the placenta normally occurs within 6 h of expulsion of the calf. Regeneration of the epithelium of the uterus postpartum is complete by 25 days postpartum, but the deeper layers of tissues are not fully restored until 6-8 weeks after calving (Sheldon et al., 2008). After parturition, the cervix reopens after 1-week postpartum and lochia is passed until 15-20 days

postpartum (Wehrend et al., 2003). Over the course of involution, lochia changes from a red-brown fluid to a more viscous yellow-white material. Uterine disease is commonly associated with *Escherichia coli*, *Arcanobacterium pyogenes*, *Fusobacterium necrophorum* and *Prevotella* species, with potential for specific virulence factors or strains of bacteria to be associated with uterine disease (Sheldon et al., 2008; LeBlanc et al., 2011).

Clinically, cows can present with puerperal metritis, clinical metritis, clinical endometritis, or subclinical endometritis. Puerperal metritis is an acute systemic illness within 10 days after parturition, being rare after the second week postpartum (Drillich et al., 2001). It is characterized by a fetid red-brown watery uterine discharge, pyrexia, reduced milk yield, dullness, anorexia, elevated heart rate, and apparent dehydration (Drillich et al., 2001; Sheldon et al., 2006; Giuliodori et al., 2013). Similarly, cows with clinical metritis present with an abnormally enlarged uterus and a purulent uterine discharge detectable in the vagina within 21 days after parturition, and without signs of systemic illness (Sheldon et al., 2006). In cases with puerperal or clinical metritis, there is an inflammation of the cavity, lining and deeper layers of the uterus (Sheldon et al., 2006, 2008). Clinical and subclinical endometritis occur 21 days or more postpartum, it is not accompanied by signs of systemic illness (LeBlanc et al., 2002) and it is characterized by superficial inflammation of the endometrium (Sheldon et al., 2006, 2008). Clinical endometritis is characterized by the presence of purulent uterine discharge in the vagina 21 days or more postpartum, or mucopurulent discharge detectable in the vagina after 26 days postpartum (Sheldon et al., 2006). Subclinical endometritis is characterized by inflammation of the endometrium in the absence of purulent material in the vagina and a cervical diameter greater

than 7.5 cm after 20 days postpartum (Gilbert et al., 1998; LeBlanc et al., 2002), and can only be diagnosed by cytology.

The terms of puerperal metritis, metritis, and endometritis are often used interchangeably, but these are different diseases based on day of diagnosis, bacterial type, and method of assessment (Sheldon et al., 2006; Potter et al., 2010). Therefore, differences in definitions of metritis and tools to assess metritis may explain variation in incidence, milk production, and reproductive status across studies.

1.3.3 Diagnosis

Metritis can be diagnosed with different techniques including rectal palpation, uterine inflammation scoring, ultrasound, radiography, and visual observations between 6 and 21 days in milk (DIM)(Andermann et al., 2007; Leutert et al., 2012). The most common method employed by producers is increased rectal temperature and manual examination of the vaginal contents, a cheap and rapid technique (Haimerl and Heuwieser, 2014). However, diagnosing metritis should not be based solely on presence of fetid, watery discharge, as there is low reliability between interobserver and intraobserver scoring systems (Sannmann and Heuwieser, 2015). Pleticha et al. (2009) attempted to create a more consistent system for evaluating vaginal discharge using a Metricheck device to consistently collect vaginal discharge samples. However, the system still requires a subjective visual scoring.

The risk factors for metritis include retained fetal membranes, calving environment, twins, dystocia, abortion, stillbirth, and diet (Gröhn et al., 1990; Correa et al., 1993; Kaneene and Miller, 1995; Sheldon et al., 2008). Metritis incidence may also vary by parity: differences in

incidence have been reported, being 34% and 56% among multiparous cows (Huzzey et al., 2007 and Sannmann et al., 2013, respectively), and greater among primiparous (61%: Sannmann et al., 2013). This suggests that confounding factors including dystocia, twins, and retained placenta may be associated with parity and differences in incidence (Benzaquen et al., 2007; Dubuc et al., 2011; Giuliodori et al., 2013). Metritis postpartum is associated with reduced milk production (Huzzey et al., 2007a; Dubuc et al., 2011; Giuliodori et al., 2013), decreased reproductive performance (LeBlanc et al., 2002; Giuliodori et al., 2013), and increased culling rate (Dubuc et al., 2011).

Little progress has been made toward the control or prevention of uterine disease. The general objective of the treatment protocols is to support and maintain innate immune function through increased dry matter intake and monitoring of NEFA and ketone bodies around parturition (LeBlanc et al., 2011). Treatment protocols are well established and reasonably effective, however, even after the resolution of the clinical signs, there is still sub-fertility (Sheldon et al., 2008).

1.4 Machine Learning

Generally, the goals in biological systems research can be summarized as 1) to formalize our understanding about the process that generates the data we observe, 2) to test an hypothesis about how the system behaves, and 3) to forecast unobserved future outcomes. Goals 1 and 2 are the emphasis of inference, helping us to understand the underlying mechanisms by creating and fitting a probability model. In contrast, goal 3 is the emphasis of prediction, which is conducted by using learning algorithms in order to find patterns with

minimal assumptions about the data-generating system. Machine learning (ML) is a group of algorithms and statistical models that computer systems use to find predictive patterns. Therefore, ML is useful in those cases where a controlled experimental design is lacking, in the presence of nonlinear interactions, when a traditional approach would be too complex, or in fluctuating environments (Bzdok et al., 2018).

The use of PDFTs in dairy farms generates large amount of data per each individual animal at high and low frequencies in fluctuating environments where nonlinear interactions are present. Given the high frequency at which changes in behavior and milk production patterns can be registered by PDFTs and analyzed by using ML systems, there is potential for developing predictive models to identify cows at higher risk of becoming clinically ill. As a consequence, earlier disease detections compared with traditional disease monitoring methods can be achieved, lowering the impact of stress and disease on animals if these are treated earlier (Weary et al., 2009; LeBlanc, 2010; Dittrich et al., 2019).

1.4.1 Machine Learning Workflow

Many methods from statistics and ML can be used for both, inference and prediction. However, the approach and workflows for inference and prediction are different. A pretty standard ML workflow involves many different steps that start as soon as we have the data available and we have defined our problem. Géron (2017) summarized a ML workflow as follows:

- Data pre-processing.
- Split the data into training and test sets.
- Model selection by fitting different models.

- Model selection by evaluation of model performance.
- Fine-tune the algorithm.
- Model assessment.

1.4.2 Data Pre-processing

Selecting the appropriate number of features (e.g., milk yield, days in milk, etc.) and knowing their importance is a critical part of a ML workflow. The process called feature engineering involves feature selection of the most useful features and feature extraction by combining existing features to create more useful ones. Common data manipulations involve handling missing values (most ML algorithms cannot work with missing values), removing duplicate records, coding categorical variables (most ML algorithms work with numbers), and feature scaling such as normalization or standardization (most ML algorithms don't perform well when the numerical variables have very different scales). It is rare to find studies in PDFT literature were the description of the steps conducted during the data pre-processing phase are included in the description of the ML workflow (Table 1.3).

1.4.3 Split Data into Training and Test Set

Prediction aims at forecasting unobserved future outcomes, but in order to achieve this goal we need 1) to train the model on sample data, and 2) to evaluate the performance of the model on future unobserved data (Bishop, 2006). Sample data is usually split into training and test sets. Typical ways to split the data is to randomly select 80% of the data as training set and 20% as test set, but other proportions such as 70/30 can be used, depending on the amount of

data available or the particular characteristics of the dataset. The training set should be representative of the cases we will generalize to once the ML system is implemented, since even in those cases when the amount of data are large, the risk of sampling bias still exists.

Performance evaluation will be conducted at two different points, being the first one after model fitting, and the second one at the very end of the workflow.

1.4.4 Fitting Different Models

Suppose we use linear regression to predict a person's weight based on a sample of individuals for whom their height and weights are known, so you can model weight as a linear function of height. In this example, the model has two parameters: the intercept and the slope, which values you can find using the least squares approach. The linear regression algorithm optimizes the parameters that make the linear model fit best to our sample data, a process also called training the model. With the estimated parameters, we could find the predicted weight of a new person whose height is known, hoping that our model will generalize well to other instances outside our sample.

A sense of how good our estimated line fits the data is given by the errors or residuals, this is, the distance between the predicted values and the actual values. It can be measured by the average of the squares of the errors, also called mean squared error (MSE), and the goal is to minimize it (Hastie et al., 2009). Common loss functions used in ML are the MSE or root mean squared error, and the mean absolute error, among others. The selection of one loss function over others will depend on a variety of factors such as presence of outliers or the choice of ML algorithm (Wang et al., 2020).

There are many different types of algorithms used in ML. Linear regression in one of them, but there are many others, and for those cases where least squares cannot be implemented to optimize the model parameters, an alternative iterative optimization approach such as gradient descent can be used. The selection of the algorithm depends on whether the outcome is known (labelled or not) in the training data. Based on this criterion, ML algorithms can be classified into unsupervised, supervised, semi-supervised, or reinforcement learning. In PDFT literature, most common ML methods fall under unsupervised and supervised learning. Therefore, we will further describe these two categories:

- *Unsupervised learning:* the training data used to train the algorithm does not include the desired solutions or labels, this is, the outcome in the training set is unlabeled. Common unsupervised algorithms are clustering algorithms (e.g., *k*-means, hierarchical cluster analysis) and dimensionality reduction (e.g., principal component analysis, independent component analysis)(Hastie et al., 2009).
- *Supervised learning:* the training data used to train the algorithm includes the desired solutions, this is, the outcome of interest in the training set is labelled. Depending on whether the outcome is continuous or categorical, problems are classified into regression or classification. The task in regression is to predict a numeric value given a set of features called predictors, while in classification is to predict the class (disease – non-disease). Being binary outcomes (healthy – sick) the most common ones in health research, ML classifiers are extensively used in the PDFT literature. Common supervised algorithms are *k*-nearest neighbors, linear regression, general linear models, logistic regression, decision trees, random forests,

neural networks, naïve Bayes, or support vector machines, among others (Hastie et al., 2009).

For disease prediction among PDFT literature, the most common ML algorithms are supervised classifiers such as *k*-nearest neighbors (Saint-Dizier and Chastant-Maillard, 2012; Shahriar et al., 2016), decision trees (Kamphuis et al., 2010b; Steensels et al., 2016; Tamura et al., 2019), random forest (Caraviello et al., 2006; Kamphuis et al., 2010b; Vanrell et al., 2014; Williams et al., 2016; Probo et al., 2018), and support vector machines (Martiskainen et al., 2009; Vanrell et al., 2014). It is not possible to know *a priori* which ML algorithm will work best for a given dataset, and the only way to find out is to fit different ones and to compare them (Wolpert, 1996).

1.4.5 Evaluation of Model Performance

Once the algorithm has been fitted to the data, we need to get a sense of how well it will generalize to new data by estimating the prediction error. We can use a proportion of the training set to fit the model and to evaluate the performance on a separate validation set. In those cases where there is no enough data to be split into different subsets, *K*-fold cross-validation can be used to split the data multiple times into train and validation sets. Specifically, the data is randomly split into equal-sized parts, and for the *k*th part, the model is fitted to the other *K*-1 pars of the data, and the prediction error of the fitted model when predicting the *k*th part of the data is calculated. The process is repeated each time until all folds have been used for both, fitting and validation, resulting in an average prediction (Hastie et al., 2009).

In the case of binary classifiers (e.g., non-diseased, diseased), there are multiple performance measures that can be used after using *K*-fold cross-validation to obtain the predicted class (e.g., non-diseased, diseased). Then, we can compare the predicted class with the actual class (0: non-cases; 1: cases), given by the labeled data (supervised learning algorithms). Based on this comparison, we can estimate the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). From these, we can further estimate sensitivity (Se; also called recall or true positive rate in ML), specificity (Sp), positive predictive value (PPV; also called precision in ML), negative predictive value (NPV), accuracy (Ac), F_1 score (the harmonic mean of precision and recall), ROC curve (Se *versus* 1 – Sp), and Precision-Recall curve (PR-curve; PPV *versus* Se). Performance metrics such as Se, Sp, PPV, and NPV depend on the threshold used to classify observations based on their class probabilities. The threshold used by default by most ML algorithms is 0.5, and if a given observation has a class probability greater than 0.5, its predicted class will be 1. Similarly, if a given observation has a class probability smaller than 0.5, its predicted class will be 0. By changing the threshold, we can increase the PPV or the Se. Selection of detection threshold is very important in those cases where the dataset in unbalanced. This is a very common situation in health applications, where the number of cases is smaller than the number of non-cases.

Common challenges in ML systems are underfitting and overfitting. Underfitting occurs when the model is too simple to learn the underlying structure of the data, resulting in inaccurate predictions. In contrast, overfitting occurs when the model is too complex relative to the amount and noisiness of the training data. As a result, the model performs well on the training data but it does not generalize well (Hastie et al., 2009). There are different ways to

control the amount of overfitting in our model such as regularization or early stopping during the learning phase of the algorithm.

1.4.6 Fine-tunning the Model

Fine-tunning in used in ML in order to achieve better model performance metrics. We can fine-tune the best selected model in multiple ways: from modifying the number of features used to fit the model, to the use of regularization to avoid overfitting. Optimize the hyper-parameters of the algorithm is also another step that is carried out during the fine-tunning process. Opposite to ordinary model parameters that are optimized during the model fitting, hyper-parameters are set by the user and are not affected by the learning process itself. Hyper-parameters must be set prior to training and the types of hyper-parameters depend on the type of algorithm used. Some examples of hyper-parameters are the number of trees in a random forest, the minimum number of sample instances at a leaf node in a decision tree, or the number of nearest neighbors used (k) in a k-nearest neighbor algorithm (Luo, 2016).

Common ways to approach hyper-parameter optimization is by providing a set of hyper-parameters on a predefined grid, and fitting multiple hyper-parameter combinations to select the one with the best performance. Two tools used in ML to search across the predefined grid are grid search and randomized search (Bergstra and Bengio, 2012). For the grid search a list of hyper-parameter values is provided and the algorithm tries all their possible combinations, using K-fold cross-validation. This process can be time consuming if the search space is large, this is, the list of the possible values for each hyperparameter is large. In contrast, randomized search

evaluates a given number of random combinations by selecting a random value for each hyper-parameter at every iteration (Bergstra and Bengio, 2012).

Among PDFT literature involving diseases postpartum it is rare to find studies where the description of how the fine-tunning was carried out is included in the ML workflow (Table 1.3).

1.4.7 Model Assessment

Once we have selected the best model based on the performance on the training and validation sets, the last step is to evaluate the performance of the ML system on new data previously unseen, or test set, by the ML algorithm. This last step is rarely seen among PDFT literature.

1.5 Research Motivation and Overview

Among PDFT literature, a relatively small percentage focuses on metabolic diseases during the transition period compared to studies on reproductive or mastitis outcomes (Rutten et al., 2013). There is a relatively large body of research which emphasis is to validate sensor measurements against visual observations and clinical findings. In contrast, fewer of these make use of ML algorithms trying to find predictive patterns in data using multiple data streams.

Monitoring a wider set of behaviors has been hypothesized to be of greater predictive value for detecting sick animals compared with more restricted set of behaviors (Mathews et al., 2016, 2017). However, despite the high number of behaviors being recorded by PDFTs, the combination of these is rarely explored and only a limited number of behavioral variables are usually included in prediction models (Saint-Dizier and Chastant-Maillard, 2018). Furthermore,

direct comparison across studies is not possible due to differences in methodologies, time windows used to aggregate sensor data, time lags, and metrics chosen. Despite the fact that classification performance is affected by the time lags chosen (Saint-Dizier and Chastant-Maillar, 2018), changes in classification performance have been ignored in PDFT literature (Carslake et al., 2021). Similarly, the study of the impact of different time windows has been poorly studied in the PDFT literature when trying to predict animal health using animal behavior, as many studies fail to establish which signal features and sampling rates are most appropriate for each behavior (Carslake et al., 2021).

Common limitations in PDFT literature are lack of control for concurrent postpartum diseases, behavioral data aggregation before and after disease diagnosis resulting in lost temporal relationships, and lack of consideration of within-same-day behavior variability due to farm scheduled activities (Huzzey et al., 2007; Stoye et al., 2012). Common limitations in the studies using ML algorithms on sensor data for postpartum disease prediction is the lack of a systematic approach to feature selection, lack of specifications regarding study approaches to minimize overfitting, and fine-tunning strategies. Furthermore, poor generalization of models in which methods to artificially balance datasets that are in nature unbalanced have been used has not been addressed.

Therefore, there is the need of a systematic approach to study algorithms performance at different time windows and time lags when inputs from multiple behaviors are included simultaneously in the model, and to combine those with other data inputs such as milk yield-related variables, without artificially balance the data.

1.5.1 Overview

Chapter 2 focused on the study of ruminating, eating, not active, active, and high activity behaviors registered by an ear-tag 3-axis accelerometer. Chapter 3 studied lying, lying bouts, steps, intake, and intake visits behaviors registered by a leg-attached 3-axis accelerometer. Hourly sensor data for each behavior and device corresponding to the 3 days before each metritis events were aggregated using time windows of 24, 12, 6, and 3 hours. Chapters 2 and 3 studied the performance of three different supervised classifiers (*k*-nearest neighbors, random forest, and support vector machines) on each individual behavior at different time windows and multiple time lags. Performance was evaluated in terms of sensitivity, specificity, positive predictive value, negative predictive value, F_1 score across multiple classification thresholds. Furthermore, area under the PR-curve and ROC curve were also evaluated. Due to the unbalanced nature of the data, a rank-based method approach was used to classify the events, and priority was given to the sensitivity, positive predictive value, and F_1 score as performance metrics. Random Forest had the greatest and most consistent performance across time windows and time lags. While 12 and 6 h time windows were best for the ear-attached accelerometer, 6 and 3 h time windows were best for the leg-attached accelerometer. Furthermore, best performances were achieved at longer time lags. Based on our findings from Chapters 2 and 3, in Chapter 4 we developed a framework for model building where features from multiple behaviors were used to fit the models using Random Forest as classifier. Furthermore, we evaluated classifier performance under nowcasting and forecasting frameworks, and we evaluated the usefulness of combining behaviors measured by two different PDFTs.

Table 1. 1: Behavioral variables registered by cow-attached precision dairy farming technologies (PDFTs), common raw data transformations (data pre-process.), and their performance metrics from validation studies for postpartum diseases.

Variables	PDFT	Frequency	Data pre-process.	Source	Performance[1]				
					r	Se (%)	Sp (%)	Ac (%)	AUC
Core Temp	DVM bolus	5 min	Mean temp/d	Bewley et al., 2008	0.65	31-41	92-95	79-92	0.54-0.77
Steps	AfiAct Pedomet.	1 h	Steps/h; Steps/2 h;	Mattachini et al., 2013	--	71-78	68-82	67-78	
	Ice Qube	15 min	Steps/d	McGowan et al., 2007	--	70-72	78-84	77-80	0.59-0.64
	TrackaCow	5 min		Wolfger et al., 2015	--	74-79	74-79	74-79	
Rumination[2]	CowManager	1 min	Min/h; % time/h;	Bikker et al., 2014	0.93	56-71	81-91	80-81	0.52-0.71
			Min/d; Units/2 h;	Borchers et al., 2016	0.69				
	HR Tag	2 h	Units/d	Burfeind et al., 2011	0.96	42-70	83-96	81-84	0.52-0.69
				Schirmann et al., 2009	0.92				
	SmartBow	15 min		Tsai, 2017	--	53-70	74-90	73-81	0.51-0.91
Lying[3]	AfiAct Pedomet.	1 h	Min/h; Min/15 min; Min/2 h;	Borchers, 2015 Mattachini et al., 2013	>0.9	71-78	68-82	67-78	
	IceQube	15 min	Min/d	Borchers, 2015 McGowan et al., 2007	>0.9	70-72	78-84	77-80	0.59-0.64
	TrackaCow	5 min		Borchers, 2015 Wolfger et al., 2015	>0.9	74-79	74-79	74-79	0.58-0.64
	SmartBow	15 min		Tsai, 2017	--	53-70	74-90	73-81	0.51-0.91
Lying bouts	AfiAct Pedomet.	1 h	Bouts/h; Bouts/2h;	Mattachini et al., 2013	--	71-78	68-82	67-78	
	IceQube	15 min	Bouts/d	McGowan et al., 2007	--	70-72	78-84	77-80	0.59-0.64
	TrackaCow	5 min		Borchers et al., 2016 Wolfger et al., 2015	>0.9	74-79	74-79	74-79	0.58-0.64
Eating	CowManager	1 min	Min/h; % time/h;	Bikker et al., 2014	0.88	56-71	81-91	80-81	0.52-0.71
			Min/d	Borchers et al., 2016	0.88				
Time at Feedbunk	TrackaCow	5 min	Min/d	Borchers et al., 2016 Wolfger et al., 2015	0.93	74-79	74-79	74-79	0.58-0.64
Feedbunk Visits	TrackaCow	5 min	Visits/d	Borchers, 2015 Chapinal et al., 2007 DeVries et al., 2003 Wolfger et al., 2015	--	74-79	74-79	74-79	0.58-0.64

Table 1.1 (Continued) Behavioral variables registered by cow-attached precision dairy farming technologies (PDFTs), common raw data transformations (data pre-process.), and their performance metrics from validation studies for postpartum diseases.

Variables	PDFT	Frequency	Data pre-process.	Source	Performance[1]				
					r	Se (%)	Sp (%)	Ac (%)	AUC
Activity	CowManager	1 min	Min/h; % time/h;	Bikker et al., 2014	0.73	56-71	81-91	80-81	0.52-0.71
	SmartBow	15 min	Min/d	Tsai, 2017		53-70	74-90	73-81	0.51-0.91
Neck activity	HR Tag	2 h	Min/2h; Min/d	Schirmann et al., 2009 Burfeind et al., 2011		42-70	83-96	81-84	0.52-0.69
Motion index	IceQube	15 min	Min/15 min	McGowan et al., 2007		70-72	78-84	77-80	0.59-0.64

[1] Pearson correlation coefficient (r) has been extracted from the sources. Performance evaluation focused on the classification of hyperketonemia, hypocalcemia, and metritis.

[2] Rumination: the point in time of regurgitation. Starts when regurgitated boluses reach the esophagus, enters the mouth, and are subsequently followed by the initiation of rhythmic chewing by the cow. Ends when rhythmic chewing ceases and the bolus is swallowed (Schirmann et al., 2009).

[3] Lying: the flank of the animal comes in contact with a surface during transition from a standing point. Standing is when the transition from a lying position to a standing position occurs and all four limbs are fully extended and perpendicular to the ground (Ledgerwood et al., 2010).

Table 1. 2: Variables registered by non-attached precision dairy farming technologies (PDFTs) at each milking, common raw data transformations (data pre-process.), and their performance metrics from validation studies.

Variables	PDFT	Data pre-process.	Source	Performance[1]			
				Se (%)	Sp (%)	Ac (%)	AUC
Milk fat	AfiLab	Daily sum; daily %; daily fat/protein ratio	Karp and Petersson-Wolfe, 2010	42-70	83-96	81-84	0.52-0.69
Milk protein	AfiLab	Daily sum; daily %; daily fat/protein ratio	Tsai, 2017	75-79	79-86	78-84	0.67-0.83
Milk yield (kg/d)	AfiMilk	Daily sum	Tsai, 2017	70-78	81-90	79-89	0.67-0.83
Milk conductivity (%)	AfiMilk	Percentage	Tsai, 2017	70-78	81-90	79-89	0.67-0.83
Body weight (kg)	AfiWeight	Mean kg/d	Tsai, 2017	28-61	80-92	81-94	0.67-0.83

[1] Performance evaluation focused on the classification of hyperketonemia, hypocalcemia, and metritis.

Table 1. 3: Use of machine learning (ML) algorithms used in precision dairy farming technology (PDFT) literature with emphasis on hyperketonemia, hypocalcemia, and metritis during postpartum.

Source	PDFT	Disease	ML Algorithm	ML Workflow[1]
Edwards and Tozer, 2004	Pedometer	Hyperketonemia Hypocalcemia	General linear model	No
Alzahal et al., 2009	Radiotelemetric bolus pH electrode	Ruminal dysfunction	General linear model	No
Bar and Soloman, 2010	HR Tag	Ruminal dysfunction	Mixed linear model	No
AlZahal et al., 2011	Radiotelemetric bolus	Ruminal dysfunction	General linear model	No
Liboreiro et al., 2015	HR Tag AfiMilk	Hyperketonemia Hypocalcemia Metritis	Logistic regression	No
Steensels et al., 2016	HR Tag	Hyperketonemia Metritis	Decision tree	Yes
Stangaferro et al., 2016a; c	HR Tag	Hyperketonemia Hypocalcemia Metritis	Logistic regression	No
Tremblay et al., 2018	AMS[2]	Hyperketonemia	k-means Principal component analysis	No
Hamilton et al., 2019	3-axis accelerometer	Ruminal dysfunction	Support vector machines	Yes
Sturm et al., 2020	SmartBow	Hyperketonemia	Nearest centroid classification Naïve Bayes classifier	Yes
Wagner et al., 2020	pH electrode	Ruminal dysfunction	k-Nearest neighbors Decision tree Multilayer perceptron Long short-term memory	Yes

[1] A machine learning (ML) workflow is considered to be included if the source describes the proportion of the data used for training and testing, methods for feature extraction, hyperparameters used, and validation.

[2] Automated Milking System (AMS) refers to milking robots where milk samples are collected in line and analyzed automatically without the need of sample handling.

2 Comparative performance analysis of three machine learning algorithms applied to sensor data in dairy cattle to predict metritis events: behaviors measured with an ear-tag accelerometer.

G. Vidal,[1] J. Sharpnack,[2] P. Pinedo,[3] I. C. Tsai,[4] A. R. Lee,[4] and B. Martínez-López[1]

[1] Center of Animal Disease Modeling and Surveillance (CADMS). Department of Medicine and Epidemiology. School of Veterinary Medicine. University of California, Davis, CA, 95616.

[2] Department of Statistics. University of California, Davis, CA 95616.

[3] Department of Animal Sciences. Colorado State University, Fort Collins, CO, 80523.

[4] Department of Animal Sciences. College of Agriculture, Food, and Environment. University of Kentucky, Lexington, KY, 40546.

2.1 Abstract

With all the sensor data currently generated at high frequency in dairy farms, there is potential for earlier diagnosis of postpartum diseases compared with traditional monitoring methodologies. Our objectives were 1) to compare the performance of three machine learning classification algorithms on the detection of behavior patterns measured by an ear-tag accelerometer (CowManager, Agis Autimatisering, Harmelen, Netherlands) associated with metritis events, 2) to determine whether farm scheduled activities have an impact on model performance, 3) to identify which behaviors yield the highest F_1 score on metritis prediction, and 4) to estimate the best time aggregation for the sensor data and how much behavioral data are necessary to obtain the highest F_1 score on metritis prediction. Data from 35 dairy cows that either did not experience any disease postpartum or were only diagnosed with metritis were retrospectively selected from a dataset containing sensor data and health information from 138 lactating cows during the first 21 days postpartum at University of Kentucky Coldstream Dairy from June 2014 to May 2017. Metritis events were created based on changes in metritis scores recorded during clinical examination. Sensor data for rumination, eating, not active, active, and high activity behaviors corresponding to the 3 days before each metritis event were aggregated every 24, 12, 6, and 3 hours, resulting in 1,386 models. All behaviors changed throughout the study period and showed distinct daily patterns. From the three algorithms, random forest had the best and most consistent performance, and scheduled activities had no impact on model performance. Furthermore, sensor data aggregated every 6 or 12 hours had the best balance between model performance and consistency of results. We concluded that the data from the first 3 days post-partum should be discarded when studying metritis, that rank-based methods

should be preferred over other methods that imply to artificially balance the data, and all five behaviors measured with CowManager were useful in predicting metritis when sensor data were aggregated every 12 or 6 hours. Findings from this study will be used to develop more complex prediction models that could identify cows at higher risk of experiencing not only metritis but other negative health outcomes.

Keywords: predictive modeling; classification algorithms; precision dairy farming; postpartum period, dairy cattle behavior

2.2 Introduction

In the 1960s and 1970s, with the development of individual cow automatic identification, information such as feeding, milk, and activity data started to be routinely collected to assist with animal management. More recently, the development of new sensor technologies has been intimately related with increasing herd sizes and labor cost that have resulted in lower ratios of farm staff to animals (Rutten et al., 2013). These new automated data collection systems generate large amounts of information in a less controlled way but at a higher frequency, increasing the volume of information that is being generated. In dairy cattle, most sensor devices record lying, feeding, and physical activity behaviors.

Sickness behavior is part of an adaptive response to infection or injury that occurs when the animal is trying to cope with a stressor. Most of the sickness behaviors are associated with depression, loss of appetite, weight loss, and pain (Tizard, 2008). However, traditional observation of sickness behavior on farms are often based upon subjective clinical evaluation, which are in turn influenced by the accumulated experience of farm personnel, with questionable consistency (Espadamala et al., 2016). The availability of new technology to automatically record behaviors allows for increased use of objective measures, and changes in behavioral patterns can be used to predict, identify, and assess health problems at the individual animal-level to prevent or mitigate clinical disease (Weary et al., 2009; LeBlanc, 2010; Dittrich et al., 2019).

Machine learning (ML) is a group of algorithms and statistical models that computer systems use to find predictive patterns. These algorithms fit models using sample data, also

called training data, and evaluate the performance of the model on new data, or test data, providing a sense of the generalizability of the models (Bishop, 2006). Therefore, ML algorithms can be used to develop predictive models to identify which cows are at higher risk of becoming clinically ill. Given the high frequency at which changes in behavior patterns can be analyzed, there is potential for earlier disease detection compared with traditional monitoring methods, lowering the impact of stress and disease on animals if these are treated earlier.

Due to metabolic challenges, diseases such as hypocalcemia, hyperketonemia, and metritis are commonly diagnosed around the transition period in 30 to 50% of dairy cattle. These health events have effects on welfare, reproductive health 1 to 9 weeks after calving, and on long term reproductive performance (LeBlanc, 2010). Despite this fact, in a review by Rutten et al. (2013), only 16% of the precision dairy farming literature were related with disease around parturition. Common limitations of these studies were lack of control for concurrent postpartum diseases, behavioral data aggregation before and after disease diagnosis resulting in lost temporal relationships, and lack of consideration of within-same-day behavior variability due to farm scheduled activities (Huzzey et al., 2007b; Stoye et al., 2012).

The objective of the present study was to compare the performance of three ML classification algorithms on the detection of behavioral patterns measured with sensor data using an ear-tag accelerometer, associated with changes in metritis score throughout the post-partum period in dairy cattle. A second goal was to identify whether farm scheduled activities had an impact on ML classification algorithm. A third goal was to determine which animal behaviors yield the highest F_1 score for metritis prediction, to estimate the best time aggregation for the sensor data, and how much behavioral data are necessary to improve metritis prediction.

Our findings will provide the foundations to develop more complex prediction models to better inform caregivers whether a medical intervention is needed.

2.3 Material and Methods

The data used in this study was part of a large study designed to quantify physiological and behavioral changes associated with mastitis, lameness, estrus, and postpartum diseases, using multiple precision dairy farming (PDF) technologies (Tsai, 2017; Lee, 2018). The larger study included data from 138 lactating cows at the University of Kentucky Coldstream Dairy (Lexington, KY, USA) that were enrolled in the study during two different periods (June 2014 to October 2015, and July 2016 to May 2017). All studies were performed with the approval of the University of Kentucky Institutional Animal Care and Use Committee (IACUC protocol number 2013-1199 and 2016-2368).

2.3.1 Population Data

From the original dataset, a total of 35 dairy cows that either did not experience any disease postpartum or were only affected by metritis were retrospectively selected. Cows were enrolled in the study after parturition and were followed for 21 days. Cows were excluded from the study if they died or were culled from the herd before 21 days in milk (DIM).

Details about farm management are described elsewhere (Tsai, 2017; Lee, 2018). Briefly, about one month before expected calving date, cows were moved from a far-off dry pen and pasture to a close-up dry pen. Cows were maintained in a fresh cow pen from parturition to 70 DIM. Subsequently, lactating cows were housed in two freestall barns. During the first study

period, one barn had 54 dual chamber waterbeds (Advanced Comfort technology, Inc., Reedsburg, WI) and the other was equipped with 54 rubber-filled mattresses, both surfaces covered with sawdust. During the second study period, both barns had compost bedded pack tilled twice daily, and bedded with sawdust as needed. Cows were provided *ad libitum* access to fresh water in each barn and lactating cows were fed the same TMR between 6:00 to 9:30 h and 12:30 to 15:00 h. The lactating diet consisted of forage, alfalfa hay, mineral and vitamin supplement, concentrate mix, whole cottonseed, and alfalfa haylage. During the second study period, feed was pushed up 22 times per day by an automated feed pusher (Lely Juno, Ley Robots, Masslius, the Netherlands). Cows were milked two times per day from 4:30 to 5:30 h and from 15:30 to 16:30 h in a double 2 X 2 tandem-milking parlor.

2.3.2 Clinical Data

Fresh cows were inspected daily after morning milking from 7:00 to 10:00 h for the first 21 days of lactation. A MetriCheck (Simero Tech Ltd, Hamilton, New Zealand) device was used to obtain a uterine discharge sample and scored on 3, 5, 7, 9, 11, 17, 19, and 21 DIM. Depending on the study period, different number of uterine discharge samples were taken between 11 and 17 DIM: during the first study period, one sample was taken on 14 DIM, while during the second study period, samples were taken on 13 and 15 DIM. A uterine discharge was evaluated on a 1 to 3 scale using a scale modified from Sheldon et al. (2006). Briefly, score 1: thick, viscous discharge, clear, opaque or red to brown in color, no odor or milk; score 2: white or yellow pus, moderate to thick discharge, milk odor; score 3: pink, red, dark red, or black watery discharge, detectable offensive odor, possibly intolerable. Cows with score $\geq$ 2 were classified as metritis

cases (Tsai, 2017; Lee, 2018). As part of the study, cows were monitored for hyperketonemia, hypocalcemia, mastitis, lameness, and retained placenta as described by Tsai (2017) and Lee (2018). Briefly, blood was collected by caudal venipuncture on 3, 7, 14, and 21 days postpartum for calcium level and non-esterified fatty acid (NEFA) determination. Beta-hydroxybutyrate (BHBA) concentration was measured with two cowside monitors: Precision Xtra (Abbott Laboratories, Chicago, IL, USA) was used on days 3, 7, 14, and 21 DIM during the first study period, while BHBCheck (PortaCheck Inc., Moorestown NJ, USA) was used on days 1, 2, 3, 4, 5, 6, 7, 10, 14, and 21 DIM during the second study period. Hypocalcemia was defined as a serum Ca level <8.6 mg/dL (Chapinal et al., 2011) and hyperketonemia was diagnosed when BHBA $\geq$ 1.2 mmol/L (Geishauser et al., 1998; McArt et al., 2012; Kaufman et al., 2016). Cows were diagnosed with clinical mastitis using the following criteria: watery, thickened, and discolored milk; milk containing blood, pus, flakes, or clots; edema, erythema; or pain in the associated quarter (Royster and Wagner, 2015) between 1 and 21 DIM by trained milkers. Furthermore, quarter milk samples were collected for somatic cell count (SCC) on days 4 $\pm$ 2 DIM and 9 $\pm$ 2 DIM. Cows with SCC $\geq$ 200,000 cells/mL in one or more quarters were considered positive for subclinical mastitis. Finally, locomotion scores were recorded on days 1, 7, 14, and 21 postpartum on a 1 to 3 scale (Schlageter-Tello et al., 2014). Retained placenta was recorded if fetal membranes were retained for > 24 hours (Sheldon et al., 2006).

For any given cow and day, a metritis event was assigned when a cow was getting or being with metritis, this is, the metritis score increased, changed from 3 to 2, or when the score remained 2 or 3, between two consecutive uterine discharge evaluations. Similarly, for any given cow and any given day, a non-metritis event was assigned when a cow was recovering from

metritis or being healthy, this is, when the metritis score decreased to 1, or when the score remained as 1, between two consecutive uterine discharge evaluations. In order to keep the time relationship between sensor measurements and clinical data, diagnosis of metritis was assigned to happen at 6:00 h on each one of the days when uterine discharge was evaluated.

2.3.3 Sensor Data and Data Pre-processing

Each cow was equipped with different PDFTs before being enrolled to allow for an adjustment period of at least two weeks. For this retrospective study, information per cow included five different behaviors measured from parturition (day 0) to 21 days postpartum with an ear-attached 3-axis accelerometer (CowManager, Agis Autimatisering, Harmelen, Netherlands) that records the number of minutes per hour for behaviors classified as ruminating, eating, not active (including both standing or lying), active, or high activity. CowManager has been previously validated by Bikker et al. (2014), and Borchers et al. (2016).

Time series sensor data consisted on the hourly measurements for each behavior i corresponding to the 3 days prior to each metritis event, assigning the time of diagnosis t at 6:00 h on each one of the days when uterine discharge was evaluated. Therefore, the 6:00 h time was used as offset for later transformations of the time series sensor data. When only sensor measurements corresponding to evening-night were used, for any given day, only sensor data from 17:00 to 3:00 h were considered, being the time of diagnosis t assigned at 17:00 h on each one of the days when uterine discharge was evaluated. The following time series data transformations were applied to both time series: one with observations for every hour, and another one containing only those corresponding to the evening-night hours.

The first time series data transformation was to remove seasonality by differencing the time series. In order to do that, we subtracted for each cow, behavior i, and hour within a 3-day period before a given metritis event, the measurement registered by the sensor in the previous 24 h from each hourly sensor measurement. The time series data for each metritis event at time t was defined by:

$$\left(x_{i,t-1}, x_{i,t-2}, \ldots, x_{i,t-n}\right)$$

where:

x_i was the differenced hourly sensor measurement for behavior i and time t, being $i \in \{rumination, eating, not\ active, active, high\ activity\}$

n was the time step within a 3-day (or 72 hours) period.

Next, we transformed the time series sensor data by aggregating the differenced hourly measurements using the mean of the time window tw_1. In order to assess classifier performance at different levels of sensor data aggregation, we used 4 different widths to compute the mean: 3, 6, 12, and 24 h. As result, the new time series data for each metritis event at time t was defined by:

$$\left(\bar{x}_{ij,t-1}, \bar{x}_{ij,t-2}, \ldots, \bar{x}_{ij,t-m}\right)$$

where:

$\bar{x}_{ij}$ was the mean sensor value for behavior i and time window tw_1 of width j, being $i \in \{rumination, eating, not\ active, active, high\ activity\}$, and $j \in \{3\ h, 6\ h, 12\ h, 24\ h\}$

m was the time step within a 3-day period. The number of time steps that could be included within a 3-day period was a function of the width j of the time window tw_1.

2.3.4 Model Fitting

We selected the number of model inputs (or features) by using a time window tw_2 of width k. In order to assess classifier performance at different widths, we used multiple values for k within a 3-day period before each metritis event. Therefore, the model inputs for each model were:

$$\left(\bar{x}_{ij,t-1}, \bar{x}_{ij,t-2}, \ldots, \bar{x}_{ij,t-k}\right)$$

where the width $k = 1, 2, \ldots, l$, and l was the number of time steps included as features within a 3-day period before a given metritis event.

The number of features in our models ranged from 1, when sensor data were aggregated with a tw_1 width j of 24 hours and tw_2 width k of 1, to 24 features when sensor data were aggregated using a tw_1 width j of 3 hours and tw_2 width k of 24, corresponding to 72 hours prior to the event.

In this paper, we evaluate the ability of 3 common supervised ML classifiers (*k*-nearest neighbors, random forest, and support vector machine). These learning algorithms can be used in classification problems using labeled data, in this case, to discriminate among 2 possible distinct patterns (metritis and non-metritis events) using sensor data from 5 animal behaviors as independent variables, also known as predictors, features, or inputs (Alpaydin, 2010).

- ***k-Nearest neighbors (k-NN).*** *k*-NN algorithm assumes that similar data points exist in close proximity, this is, are close to each other. The algorithm estimates the closeness by calculating the Euclidean distance for each data point to the rest of the data points, sorts the distances from smallest to largest, and picks the first *k* entries

and their labels (metritis, non-metritis events), returning the mode of the *k* labels (Fix and Hodges, 1951; Dasarathy, 1991; Hastie et al., 2009).

- ***Random forest (RF).*** Random Forest is a model made up from many decision trees. A decision tree can be seen as a flowchart of questions asked about the data, eventually leading to a predicted class with the greatest reduction in Gini Impurity. This means that the decision tree tries to form nodes or sets of data points that are as pure as possible, containing a high proportion of data points from only one class (metritis, non-metritis events). Therefore, the Gini Impurity is the probability that a randomly chosen sample in a node would be correctly labeled if it was labeled by the distribution of samples in the node. In a decision tree, for each level of the tree, the weighted average Gini Impurity decreases as we approach the terminal nodes, also called leaf nodes, and the class is the majority classification, or prediction, for the data points in the node. The number of levels in a decision tree can be controlled by limiting the maximum depth of the tree. When the depth has no limit, the tree is allowed to create as many levels as necessary in order to classify all the points, overfitting the data by growing until it has one leaf node for every single observation. To deal with overfitting, we can use hundreds or thousands of decision trees to form a forest, and the final prediction then becomes the average prediction from all trees in the ensemble. The model is random because uses 1) random sampling of data points, with or without bootstrapping, to generate each decision tree; and 2) splitting nodes based on a limited number of the features (Breiman, 2001; Hastie et al., 2009).

- *Support vector machines (SVM).* In cases where two classes can be linearly separated, we can divide the data points with a line into two regions labeled according to the classification, this is, metritis and non-metritis events. Examples of these methods are linear regression, linear discriminant analysis, and logistic regression. For those cases in which there are more than two classes, or when the different classes overlap, we can use SVM. In this case, instead of dividing the data with a line, we will need to estimate the optimal hyperplane, also called decision hyperplane, that separates the different classes as well as possible while maximizing the distance, also called margin, to the closest point from either class, also called support vectors. For those cases in which classes are not linearly separated, a kernel function can be used where data points can be mapped to a transformed version of the feature space so data can be then linearly separated. To deal with the overlap while maximizing the margin, we need to allow for some points to be on the wrong side of the margin. This is controlled by the cost parameter C, which allows data points to fall off the margin, controlling the tradeoff between the misclassifications and width of the margin (Vapnick, 1995; Hastie et al., 2009).

For each one of the three ML classifiers, one model was fitted for each combination of behavior i, time window tw_1 width j, and time window tw_2 of width k.

2.3.5 Model Performance

Due to limitations in the amount of data available, we used fivefold cross-validation (5-FCV) to set aside a validation set and use it to assess the performance of the prediction model.

Specifically, for any given model, the data were randomly split into 5 equal-sized parts where 4/5ths were used to fit the model, whereas the 1/5th was used to calculate the prediction error of the fitted model. This process was repeated each time until all 5 folds had been used for both, fitting the model and validation, resulting in an average prediction error. First, algorithm default values for the different hyperparameters were used to estimate classification performance using 5-FCV. Next, Grid Search (GS) was used as strategy to optimize the classifier, except for random forest, where GS was performed after Randomized Search (RS) in order to reduce the grid search so computing time was manageable. Optimal parameters that were found to allow for best mean cross-validation accuracy were used to train the final model (Table 2.1). After optimization, for each model that was fitted using 5-FCV, the prediction class probability for each health event of being classified as metritis was obtained and ranked from highest to lowest. To estimate the performance of each model, highest 20, 30, and 40% class probabilities were used as different cut-off points. For each cut-off point, classification performance was evaluated using estimates of sensitivity (Se or recall), specificity (Sp), positive predictive value (PPV or precision), negative predictive value (NPV), accuracy (Ac), F_1 score, the area under the curve (AUC) for the receiver operating characteristic (ROC) curve and Precision Recall (PR)-curves. Sensitivity is estimated as the ratio of correctly predicted positive observations to all observations in the actual class (metritis event). Specificity is estimated as the ratio of correctly predicted negative observations to all observations in the actual class (non-metritis event). Positive predictive value is the ratio of correctly predicted positive observations to all predicted positive observations. Similarly, negative predicted value is the ratio of correctly predicted negative observations to all predicted negative observations. Accuracy is the ratio of correct predictions to all number of observations

(Hogeveen et al., 2010). F_1 score is the weighted average of PPV and Se. This score takes both false positives and false negatives into account, and it is more useful than accuracy in situations where the distribution of the observations in each class is unbalanced. F_1 score was computed as $(1 + b^2)*(PPV * Se) / ((b^2 * PPV) + Se)$, where b = 1 (Saito and Rehmsmeier, 2015).

Classifier implementations were taken from the open source Python library scikit-learn (Pedregosa et al., 2011). The feature extraction and the optimization of the classifier parameters were implemented using Python programming language, version 2.7 (Python Software Foundation, http://www.python.org). Plots were done using ggplot2 library (Wickham, 2009), using R open-source statistical software (R Core Team, 2017).

2.4 Results

A total of 35 dairy cows (Jersey = 20; Holstein = 15; primiparous = 17; multiparous = 18) were retrospectively selected from the original dataset (n = 138) containing clinical and sensor data from parturition to 21 DIM. Average $\pm$ SD milk yield was 36.1 kg. $\pm$ 15.6. Of the 35 cows selected, 13 did not have any metritis event during the study period, while 22 were diagnosed at least once with metritis (score $\geq$2), occurring on average at 12 DIM (12.02 $\pm$ 4.72 DIM). Among these, 2 cows had retained fetal membranes and were kept for data analysis. None of the selected animals had hyperketonemia, mastitis, or hypocalcemia. The proportion of metritis events for primiparous and multiparous cows were 20% and 23%, respectively. Based on the changes of metritis score between two consecutive evaluations, 239 health events were created, and of those, 188 were in the non-metritis event class, while 51 were in the metritis event class. Our data set was unbalanced given the larger number of observations from the non-metritis

events class to the smaller number of observations from the metritis events class. For each one of the behaviors measured, we obtained 11,530 records.

Cows showed high variability in their behaviors during the study period, especially regarding mean time spent not active, rumination, and eating (18.97 $\pm$ 13.81 min/h, 24.07 $\pm$ 13.27 min/h, 9.41 $\pm$ 11.43 min/h, respectively; Table 2.2). Furthermore, the distribution for the behavior variables eating, active, and high activity were right-skewed, with differences in the mean values of high activity behavior between primiparous and multiparous cows when the time was categorized into milking, morning, and evening-night. Overall, differences between primiparous and multiparous seemed to be lesser during milking times (Figure 2.1).

2.4.1 Changes in Behavior by Days in Milk

There were changes across all behaviors from parturition to 21 DIM, with significant variation in the first 3 days post-partum for rumination, eating, high activity, and active behaviors. During the first 3 days post-partum, there was an increase in the amount of time cows performed rumination behavior while there was a decrease in the amount of time performed eating, high activity, and active behaviors. Overall, not active and active behaviors had a downward trend while rumination and eating had an upward trend during the study period. When behaviors were stratified by parity, primiparous spent more time performing high activity behavior and less time performing not active behavior than multiparous between 5 and 15 DIM. Overall, primiparous cows spent greater time performing eating, active, and high activity behaviors than multiparous cows, although differences were not always significant (Figure 2.2).

2.4.2 Changes in Behavior by the Time of the Day

When looking at the variation of the behaviors throughout the day, cows showed inverse trends regarding rumination compared to eating behaviors. Animals spent greater time ruminating at night, while eating behavior steadily increased from 4:00 h until it reached a first peak at 10:00 h, with a second peak around 18:00 h. The lowest observed time spent eating occurred right before 4:00 h and around 15:00 h. Active and high activity behaviors showed similar trends with respect to each other, and inverse trends when compared with not active behavior. Active and high activity behaviors peaked at 19:00 h, with increased activity levels of smaller magnitude around 5:00 h and 10:00 h. In contrast, cow behavior classified as not active was more prevalent from 0:00 to 6:00 h. When behaviors were stratified by parity, despite having similar trends, primiparous cows spent more time performing high activity behavior compared with multiparous cows on any given day, while there were no differences by parity for other behaviors measured with CowManager (Figure 2.3).

2.4.3 Changes in Behavior by Time of the Day Stratified by Days in Milk

To further explore the changes of the different behaviors across the study period, we also looked into the variation throughout the day when DIM was categorized into 3 distinct periods: convalescent (from parturition to 3 DIM), first week (4 to 7 DIM), second week (8 to 14 DIM), and third week (15 to 21 DIM). The amount of time performing each behavior changed across the different periods of the study, being the difference between the convalescent period and the third week the most evident. The amount of time performing not active, active, and high activity

behaviors was greater during the first 3 DIM, while cows spent less time performing rumination behavior during the same time period (Figure 2.4).

2.4.4 Classifier Performance

In order to estimate the performance of each one of the classifiers across all time windows tw_2 within a 72 hours period, a total of 45 models with different number of features were fitted for each combination of behavior and classifier: 3 models for sensor data aggregated every 24 hours (e.g. $j = 24h$ and $l = 1, 2, 3$), 6 models for sensor data aggregated every 12 hours (e.g. $j = 12h$ and $l = 1, 2, \dots, 6$), 12 models for sensor data aggregated every 6 hours, and 24 models for sensor data aggregated every 3 hours. A total of 45 models were fitted for each combination of classifier and behavior for all day time series sensor data. Model fitting was repeated using evening-night sensor data only. For the time series containing sensor measurements for the evening-night hours only, there was no difference in the data points used to make the 24- and 12-hour data aggregations. As a consequence, a total of 21 modes were fitted for each combination of classifier and behavior for evening-night only time series sensor data. The process of model fitting was repeated a third time for those variables that showed differences by parity and by hour. Hence, stratified models by parity were fitted for high activity and not active behaviors (all day and evening-night only time series data). A total of 1,386 models were fitted to account for differences by parity, by width j of the time window tw_1 used to aggregate the sensor data, by the width k of the time window tw_2, and by time of the day (all day, evening-night time series data). 5-FCV F_1 scores at the 20% cut-off was used to compare across different models. For all classifiers, higher F_1 scores were obtained when the time series

including sensor data for each hour was used (Figure 2.3). Figure 2.4 shows an overview of the distribution of F_1 scores at the 20% cut-off at different times before the health event, stratified by the different levels of sensor data aggregations, when sensor data for all day were used. Random forest had the highest and most consistent F_1 scores across multiple levels of data aggregations, followed by *k*-NN and SVM. Metrics performance from all the models can be found in a data repository (Vidal et al.).

To select amongst the best models, we focused on those that were in the upper quartile of the F_1 score distribution at the 20% cut-off within each classifier, when sensor data for all day were used. For RF, the upper quartile for F_1 score values at the 20% cut-off were between 94.74% and 98.76%, while the upper quartile for *k*-NN was between 41.97% and 58.82%. In contrast, the top 25% values for SVM F_1 scores were between 25.32% and 66.67% (Figure 2.5). Figure 2.6 shows the best models considered so far at the different levels of time series sensor data aggregations and number of time steps before the health event. Our results showed that RF had the best performance at any level of time aggregation, and the differences in performance between *k*-NN and SVM decreased as the sensor data became less aggregated (e.g., 3h time window). When data were aggregated using 24 hour time window, active, eating, and high activity were the predominant behaviors amongst the best models. All five behaviors were present amongst the best models when data were aggregated either using 12 or 6 hour time windows, although *k*-NN and SVM yielded less consistent results. When data were aggregated using 3 hour time window, not active, rumination and eating were the most prevalent behaviors. Models stratified by parity were also fitted for high activity and not active behaviors. While not

active behavior performed better when sensor data were not stratified, high activity performed slightly better when sensor data were stratified by parity, regardless of classifier (Figure 2.7).

2.4.5 Best Classifier, Time Window, and Time Lag

In our study, the best balance between number of behaviors, and consistency regarding number of times before an event a given behavior ranked amongst the best models was found when RF was used with data aggregated with 6 or 12 hour time windows. When data were aggregated using 12 hour time windows, best models were found for the previous 36 to 72 hours before event (time steps before event = 3, 4, 5, and 6). When data were aggregated using 6 hour time window, most consistent results were found between 42 to 72 hours before event (time steps before event = 7 to 12). Table 2.3 and 2.4 show the performance metrics for the selected best models at two different cut-off percentages. Sensitivities were slightly higher when data were aggregated using 12 hour time window and increased as we got closer to the event. Using the estimated predicted class probabilities, we compared the metritis events identified by RF with the clinical data. We found that the missed events occurred in 2 or 3 cows, depending on the level of sensor data aggregation and number of time steps before event. Among these, the majority had had multiple metritis events throughout the study, and those missed occurred when the cow had previously been identified with metritis at least once and was, most likely, undergoing medical therapy. The other situation where an event was missed occurred in a cow with just one metritis event on day 11 postpartum. All cows starting with severe metritis (score 3) were correctly identified.

2.5 Discussion

In this paper, we compared the performance of three different classification ML algorithms on five different behavior variables. Our goal was to assess the ability of the classification algorithms to identify patterns in the sensor data that may be associated with changes in metritis score during the first 21 days post-partum. To preserve the time structure of the sensor data, metritis events were created and sensor data from the 3 days prior to an event were aggregated using 4 different time windows. To our knowledge, this is the first paper studying the potential use of sensor data for disease association between sensor and metritis data, where time structure has been kept intact while controlling for other diseases post-partum.

In this study, the dataset was unbalanced, which is a common problem when farm management is appropriate and the disease to non-disease ratios are small. Most ML algorithms focus on minimizing classification errors, favoring the majority class and making the minority class harder to predict. Different methods are being used to approach unbalance data and these can be broadly classified as: 1) under-sampling, or to leave out data from the majority class to match the number of samples coming from each class; 2) over-sampling, creating synthetic samples that belong to the minority class; 3) resample, using bootstrap with weights so new class sizes come out equal; and 4) ranking-based approach, where observations are ranked based on their classification probabilities. Methods 1 through 3 do artificially balance the data, however, these approaches don't necessarily fix the problem. Sometimes, the patterns for a given class may be harder to learn and balancing the data may decrease the presence of such class, resulting in overall worse performance (Lacroix et al., 1997). Algorithm performance also

depends on the behaviors, with activity states being easier to classify by RF compared to resting behaviors (Williams et al., 2016). In contrast, a rank-based approach like the one used in this study has a number of attractive features: does not create false assumptions regarding the true population proportion among classes, does not underestimate the leverage of outliers on the estimates, and does not imply the removal of data that has a cost to collect, store, and process. We expected ranking-based approaches to yield generalizable results in those cases where the ultimate goal is to build predictive models. Once models are fitted, different metrics can be used to quantify classifier performance. Accuracy and AUC for the ROC are appropriate metrics when the minimization of classification error is the priority and datasets are balanced, and these two are widely used in precision dairy farming and disease postpartum studies. However, with unbalanced datasets, metrics such as F_1 score and PR-curve AUC may be preferred (Davis and Goadrich, 2006; Saito and Rehmsmeier, 2015). In our study, non-disease events were the majority class, and as a consequence, Sp was expected to be high regardless of how good or bad the classifier is with this type of data. Since our priority was to improve the Se and PPV of our models, we prioritized the F_1 score, the harmonic mean of PPV and Se with equal weights for both. Is worth noting that, in order to account for the different costs different misclassification errors may have, weights for PPV and Se can be modified adapting the formula as follows: $F_b = (1 + b^2) * (PPV * Se) / (b^2 * PPV) + Se$. To account for the trade-off between Se, Sp, PPV, and NPV, we explored different classification probabilities cut-offs. We found that when the threshold value was decreased from the top 20% to the top 40% most likely of being a metritis event, the number of false positives increased, which was in agreement with other studies (Shahriar et al., 2016).

In this study, the behaviors changed according to the DIM of the animal. Our findings showed that most behaviors have a trend during the first 3 days post-partum that differs from the trend observed during the rest of the study period. Activity and high activity had an overall downward trend, while eating declined during the first 3 DIM with an overall upward trend after the convalescent period. In contrast, rumination behavior had an upward trend that was steeper during the first 3 DIM. Increased rumination and eating time have been observed in the 4 -8 hours after parturition in other studies (Schirmann et al., 2013; Pahl et al., 2014). Furthermore, feeding behavior has been found to decrease by 35% over the 2 weeks before calving and to increase by 99% over the 3 weeks following parturition (Urton et al., 2005). Not only did behaviors change according to DIM, but also according to the time of the day. When behaviors were observed in a 24 hour period, there were differences regarding the type of behaviors animals would perform. Rumination was more prevalent during nighttime and had opposite trend compared with eating, a behavior that decreased before A.M. milking when the food available was the oldest sitting feed. Eating behavior steadily increased between A.M. milking and 10 h and dropped again during the hottest part of the day around 15 h. Not active behavior was also more prevalent during nighttime, when cows spent more time lying and was the coolest part of the day, and had opposite trend to active and high activity, which increased during milking and feeding times. Our findings are in agreement with previous studies, where rumination and eating time have been found to have opposite trends, with rumination time being more prevalent during nighttime (Soriani et al., 2013). Despite the differences in behavior throughout the day, we found that farm scheduled activities had no effect on the performance of each classifier and, in fact, performance was slightly better when all sensor data regardless of

time of the day were used, possible due to a maximization of the sample size. Nevertheless, depending on the goal of the study, there are some behavior data that may not be worth considering given certain times of the day, particularly during milking times, where behaviors such as lying cannot be performed. If the goal is to detect cows in heat, removing sensor data from the milking will reduce the number of false positives (Shahriar et al., 2016). If the goal in removing data is to reduce noise, then other strategies such as stratification by DIM may be more relevant. Using all sensor data or just data from certain parts of the day should be studied on a case basis, taking into account the number of animals in a given farm with sensor data, and the goal of the predictive model.

To better understand the dynamics of cow behavior throughout the study period, we looked at the different behaviors in a 24 hour period when DIM is categorized. We found that behaviors progressively evolved as we got closer to the end of the study, so the convalescence period was very different from what was observed during the last week of the study. During second and third week, behaviors became more similar to the crude data, having a more marked pattern if we consider the scheduled farm activities. Based on our findings, we conclude that sensor data from the first 3 days post-partum should be disregarded as, most likely, adds noise to the data given current management practices around calving (frequent movements between pens, frequent health checked, first time in milking parlor, etc.). In our study, the first and second metritis score evaluation happened at 3 and 5 DIM, and metritis events were only found after 5 DIM. Since the metritis (or non-metritis) events were created based on the change (or lack thereof) in the metritis score, the earliest an event can occur is at 5 DIM and therefore, it is safe to assume that the convalescent period is not affecting classifier performance. Animal

behavior can also change based on different animal-level factors. In this study, we found differences by parity for high activity behavior that started to appear between 4 and 7 DIM. Our finding is in agreement with what has been found in grazing cattle (Williams et al., 2016). These findings may be explained by social dominance dynamics between primiparous and older cows (Sepúlveda-Varas et al., 2014).

One of our goals in this study was to determine, among three different ML classifiers, which one performed best using F_1 score at the 20% cut-off as criterion to select amongst all the models. Based on the F_1 score distribution, RF had higher F_1 score values overall and better performance than *k*-NN or SVM classifiers. Furthermore, even though *k*-NN and SVM had some models with high F_1 scores, RF had consistently higher F_1 scores across all the different time aggregations. In this study, amongst those models with best performance, *k*-NN achieved an F_1 score with values around 50%, while SVM F_1 score values were from 25 – 65%. Regardless of classifier, fewer models ranked among the best when sensor data were aggregated using 24 hour *versus* 3 hour time windows (Figure 2.4). *k*-NN is a classifier that is simple and easy to implement. It has been previously used in heat detection studies, with accuracy of 82 – 100%, Se over 80%, and Sp between 90 – 100% (Saint-Dizier and Chastant-Maillard, 2012; Shahriar et al., 2016). In contrast, in this study, accuracy, Se and Sp were around 79%, 49%, and 87%, respectively, yielding suboptimal results when *k*-NN classifier was used with our data. The differences observed between this study and others are most likely due to the unbalance dataset used in this study compared with heat detection studies, where a higher proportion of animals is expected to have a positive outcome.

Of the three classifiers studied, SVM had the poorest performance, having the lowest values for Se, Sp, PPV, NPV, and Ac. Interestingly, SVM was the only classifier whose F_1 score improved as sensor data were using 3 hour time windows instead of 24, 12, or 6 hours, increasing from 25% to F_1 scores around 60%. This has also been observed in previous studies, where SVM has been reported to have better performance with shorter time aggregations (Vanrell et al., 2014). In dairy cattle, SVM classifier has been mainly used to test the ability of accelerometers to identify cows in heat (Vanrell et al., 2014) and in sensor calibration studies (Martiskainen et al., 2009). Due to the differences in the nature of these studies, results regarding classifier performance are not comparable across studies. Among all the classifiers, RF had the best performance, with Se values around 92%, and Se, PPV, NPV, and Ac higher than 98%. As described above, RF is based on decision trees, a classification method that has been used in the precision dairy farming with great success to study grazing cattle behavior (Williams et al., 2016), to predict fertility and improve heat detection in dairy cows (Caraviello et al., 2006; Vanrell et al., 2014), to predict mastitis (Kamphuis et al., 2010a), or to understand relationships between metabolic diseases postpartum and culling risk (Probo et al., 2018). Compared to other classifiers, RF can handle large data sets with high number of features, but interpretation is less intuitive when trying to understand the relationship existing in the input data. Nevertheless, we were able to link the classification probabilities with the clinical data, finding that, among those metritis events that were misclassified as false negatives by the model at the 20% cut-off, the majority were cows that were undergoing medical treatment.

In our models, best results were obtained using RF for activity (high activity in primiparous, active, eating, and rumination) as well as resting states (not active). Activity is a sign

of estrus in dairy cattle and, consequently, increased physical activity has been studied to improve heat detection (Firk et al., 2002). Activity can be measured with pedometers or accelerometers that transform the acceleration into angles. When an accelerometer is attached to the ear, changes in angles are interpreted as rumination, eating, or activity (Saint-Dizier and Chastant-Maillard, 2012). However, decreased activity is also a sign of sickness behavior whose goal is to conserve energy and, therefore, it could be used to detect disease. Changes in activity have been observed in cows that have suffered metabolic or digestive disorders postpartum (Edwards and Tozer, 2004). Most importantly, other authors that have found decreased activity before and even beyond metritis diagnosis (Liboreiro et al., 2015; Stangaferro et al., 2016a; Steensels et al., 2017). Feeding behavior is an activity state that has been found to decrease in cows with metritis, with a Se ranging between 71 – 89%, and Sp between 62 – 77% when multivariable logistic regression is used as classifier (Urton et al., 2005). In our study, we found higher Se and Sp (90 – 95.12%, 100%, respectively), showing that RF is a better than logistic regression for this type of data, where a linear separation between classes is not possible, indicating that decision tree methods may be more appropriate.

In this study, rumination showed high variability; the mean rumination time was 577.68 minutes per day (mean 24.07 minutes per hour $\pm$ 13.27 SD), and is within the range found by others (Zebeli et al., 2006; White et al., 2017). This high variability can be attributed to animal variability, milk yield, dry matter intake, chemical and physical characteristics of the diet, and to differences in the measuring technique (Beauchemin, 2018). Rumination is important for an efficient digestive function that tends to occur during nighttime in association with lying and in opposition to eating behaviors. Rumination has been found to decrease before metritis

(Liboreiro et al., 2015; Stangaferro et al., 2016a; Steensels et al., 2017) as well as during estrus, subclinical acidosis, parturition, disease, and acute stress (Herskin et al., 2004; Beauchemin, 2018). In this study, Se was 90 – 95%, Sp was 98.74 – 100%, PPV was 95 - 100%, NPV was 97.24 – 97.97%, and Ac was 97 – 90% for rumination when the 20% cut-off was used. Our findings were higher than those found by (Paudyal et al., 2018), with reported value ranges for PPV and Se of 40 - 60% and 52 – 63%, respectively. Similarly, model performance in this study was superior to that found by (Stangaferro et al., 2016a), with reported values for Se, Sp, PPV, NPV, and accuracy of 59%, 97.6%, 58.3%, 97.7%, and 95.6%, respectively. Both of these studies had unbalanced datasets, with fewer animals in the metritis class, however, none of these used rank-based methods for the classification probabilities, nor they reported their different metrics at different cut-offs. Opposite to activity state are behaviors that can be considered as resting state. Behavior classified as not active can be interpreted as standing or lying down, because the way the CowManager works. Previous studies have reported increased time standing inactive before calving and before metritis diagnosis (Patbandha et al., 2012).

In the precision dairy farming field, multiple levels of time aggregations have been used for time series data: from seconds (Vanrell et al., 2014), to days (Paudyal et al., 2018), to weeks (Tsai, 2017; Lee, 2018). To our knowledge, this is the first study where throughout comparisons have been made across different time windows. We found that the number of models that ranked amongst the best ones changed based on the different time windows: fewer models were found when sensor data were aggregated using 24 hour time windows compared with the 3 hour time window. Furthermore, when data were aggregated using 3 hour time windows, some behaviors were less consistent in their results: some behaviors ranked amongst the best

model some time before the event but not in a consistent manner. We found that the best balance between number of behaviors ranking amongst the best models and consistency of results were obtained when sensor data were aggregated using 12 or 6 hour time windows. We hypothesize that, when data are aggregated every 24 hours, fewer features are included as inputs and disease patterns in the sensor data may be masked by noise, as data non relevant with the sickness behavior are also included in the data sample (i.e., activity level, or eating time during milking times). As result, fewer models yield high Se and PPV estimates. In contrast, aggregating sensor data using 6 or 12 hour time windows include higher number of model inputs, allowing for more opportunities to detect patterns associated with cows at risk of metritis. When sensor data are aggregated every 3 hour time windows, the classifier may be finding patterns associated with disease that, in the bigger picture, do not translate well into behavioral changes since results are less consistent. Besides, the computing cost of analyzing a higher number of features such as those generated when data is aggregated every 3 hours does not seem to translate into a better performance overall. In this study, we also found that all five behaviors measured with CowManager device can be used to detect cow at risk of a metritis event using sensor data from 1.5 to 3 days before the event as model inputs. This is in agreement with what has been found by other authors: rumination time has been found to change 2 - 3 days prior to metritis diagnosis (Steensels et al., 2017; Paudyal et al., 2018) while activity has been found to change 2 days before diagnosis of metritis as well as metabolic and digestive problems (Edwards and Tozer, 2004; Steensels et al., 2017).

Limitations remain with current prediction models regarding how to deal with cases of more than one illness and how to detect one illness without excluding the others from analyses,

how classifier performance can change by adding multiple behaviors or devices at the same time, without artificially balancing the data, and how model performance translate in commercial farms in terms of Se and PPV when interactions between behavior and response to therapy occur. Future studies to address these aspects are highly needed.

2.6 Conclusions

The findings of this study have a number of practical implications. Our results indicate that data from the first 3 DIM should be studied as a complete separate period of time when studying metritis events. The second major finding was that rank-based methods for model fitting yields superior results to those studies where data were artificially balanced. Therefore, rank-based methods should be preferred when developing predictive models that will be implemented in the future. Lastly, we found that activity data, eating, and rumination time can be used to predict metritis events when sensor data is aggregated using either 12 or 6 hour time windows.

2.7 Acknowledgements

The authors would like to thank the University of Kentucky Coldstream dairy staff, and to all the students who helped with fresh cow exam and data collection. We would also like to thank Jeffrey Bewley for facilitating data sharing. The work was partially supported by NSF award IIS-BigData-AI-1838207. JS is partially supported by NSF DMS 1712996.

Table 2. 1: Hyperparameter values used for optimization of k-nearest neighbors (k-NN), random forest (RF), and support vector machines (SVM) classification algorithms used on behavioral variables measured with CowManager.

Classifier	Hyperparameter	Randomized Search	Grid Search	Models Used[1]	Optimum Value
k-NN	*k*[2]	N/A	1 to 15	tw_1 j = 24h, tw_2 k = 1, 2, 3	8
RF	Bootstrap[3]	True, False	True	tw_1 j = 24h, tw_2 k = 2	True
	Max. depth[4]	10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, None	10, 20, 30, 40	tw_1 j = 3h, tw_2 k = 24	10
	Max. features[5]	'auto', 'sqrt'	'sqrt'		'sqrt'
	Min. samples leaf[6]	1, 2, 4	1, 2, 3, 4		2
	Min. samples split[7]	2, 5, 10	2, 3, 4, 5, 6		4
	Number of estimators[8]	100, 200, 300, 400, 500, 600, 700, 800, 900, 1000	200, 400, 600, 800		300
SVM	Kernel[9]	N/A	Linear, rbf, poly, sigmoid	tw_1 j = 24h, tw_2 k = 2	Linear
	C[10]	N/A	0.01, 0.1, 1, 10, 100, 1000	tw_1 j = 3h, tw_2 k = 24	0.01
	Degree[11]	N/A	2, 3		2
	Gamma[12]	N/A	'auto', 0.01, 0.1, 1, 10, 100		'auto'

[1] All models used to find the optimum values for each hyperparameter included rumination data, and for each combination of time window j and number of time steps k, two different time series were used: all sensor data regardless of the time of the day (all day time series), and sensor values corresponding to the evening-night hours only (evening-night time series).

[2] *k*: number of neighbors.

[3] Bootstrap: method for sampling data points (with or without replacement).

[4] Max. depth: maximum number of levels in each decision tree to control for overfitting.

[5] Max. features: maximum number of features considered for splitting a node.

[6] Min. samples leaf: minimum number of data points allowed in a leaf node.

[7] Min. samples split: minimum number of data points in a node before the node is split.

[8] Number of estimators: number of trees in the forest.

[9] Kernel: type of kernel used to map the data to a different space where a linear hyperplane can be used.

[10] C: cost parameter to control the tradeoff between the misclassifications and width of the margin.

[11] Degree: degree of the polynomial used when kernel = 'poly'.

[12] Gamma: defines how far the influence of a single data point reaches and configures the sensitivity to differences in the data. When gamma is large, the radius of the area of influence only includes the support vector itself, and no amount of regularization with C will be able to prevent overfitting.

Table 2. 2: Descriptive statistics for the five behavior variables measured with an ear-attached 3-axis accelerometer (CowManager, Agis Autimatisering, Harmelen, Netherlands).

		Time Aggregation[1]				Time of the Day[2]		
	Raw Data	3 hrs.	6 hrs.	12 hrs.	24 hrs.	Milking	Morning	Evening-Night
Rumination (minutes per hr.)								
n	11,530	3,736	1,967	1,014	522	2,083	4,684	4,763
mean	24.07	0.38	0.41	0.53	0.50	24.16	21.84	26.22
std	13.27	8.80	7.05	4.90	3.34	11.83	12.98	13.77
min	0	-38	-38	-23	-12.57	0	0	0
25%	14	-5	-3.33	-2.33	-1.49	16	11	16
50%	24	0.33	0.333	0.5	0.42	24	22	27
75%	33	5.67	4.2	3.16	2.32	32	31	36
max	60	43	43	43	18.5	60	60	60
Eating (minutes per hr.)								
n	11,530	3,736	1,967	1,014	522	2,083	4,684	4,763
mean	9.41	-0.04	-0.07	-0.09	-0.09	7.02	11.30	8.60
std	11.43	6.87	5.29	3.42	2.41	8.28	12.40	11.33
min	0	-41	-41	-41	-20.17	0	0	0
25%	1	-3.33	-2.5	-1.70	-1.42	1	1	0
50%	5	0	0	0	0.09	4	7	3
75%	14	3.33	2.45	1.70	1.36	10	17	13
max	60	44	44	16	7	52	60	60
Not Active (minutes per hr.)								
n	11,530	3,736	1,967	1,014	522	2,083	4,684	4,763
mean	18.97	-0.17	-0.17	-0.25	-0.18	20.47	18.96	18.32
std	13.81	10.22	8.20	5.78	4.29	12.09	14.30	13.98
min	0	-55	-55	-48	-20.17	0	0	0
25%	8	-6	-4.5	-3.17	-2.32	11	7	7
50%	17	-0.33	-0.17	-0.29	-0.14	19	17	16
75%	28	6	4	2.67	1.94	29	28	28
max	60	57	57	36.58	34	60	60	60
Active (minutes per hr.)								
n	11,530	3,736	1,967	1,014	522	2,083	4,684	4,763
mean	3.98	-0.08	-0.08	-0.08	-0.09	4.46	4.01	3.74
std	4.09	2.72	2.10	1.48	1.21	3.66	4.09	4.25
min	0	-19.33	-12	-9.33	-9.33	0	0	0
25%	1	-1.33	-1.17	-0.83	-0.67	2	1	1
50%	3	0	0	0	0	4	3	2
75%	6	1.33	1	0.75	0.5	6	6	5
max	46	18.67	12	6.5	5.06	34	44	46

Table 2.2 (Continued): Descriptive statistics for the five behavior variables measured with an ear-attached 3-axis accelerometer (CowManager, Agis Autimatisering, Harmelen, Netherlands).

		Time Aggregation[1]				Time of the Day[2]		
	Raw Data	3 hrs.	6 hrs.	12 hrs.	24 hrs.	Milking	Morning	Evening-Night
High Activity (minutes per hr.)								
n	11,530	3,736	1,967	1,014	522	2,083	4,684	4,763
mean	3.92	-0.09	-0.10	-0.11	-0.13	4.28	4.25	3.44
std	4.93	3.03	2.27	1.68	1.34	4.83	5.10	4.77
min	0	-16.67	-13.5	-8.22	-8.17	0	0	0
25%	0	-1.33	-1.17	-0.85	-0.79	1	1	0
50%	2	0	0	0	-0.07	3	3	2
75%	6	1.33	1	0.74	0.58	6	6	5
max	40	19	13.67	8.42	5.46	37	40	38

[1] Time aggregation was done after differentiation of the raw sensor data.

[2] Time of the day: for data analysis purposes, milking is from 4:00 to 5:59 h and from 15:00 to 16:59 h; morning is from 6:00 to 14:59 h; evening-night is from 17:00 to 3:59 h in the following day.

Table 2. 3: Results from models performance (%) where random forest (RF) classifier was used on sensor data from an ear-attached 3-axis accelerometer (CowManager, Agis Autimatisering, Harmelen, Netherlands) device from all day using a 12 hour time window. Different cut-off values were chosen based on the highest classification probabilities

Behavior	Time Lag *k*	Sample Size		20% cut-off						30% cut-off					
		Metritis	Non-metritis	Se	Sp	PPV	NPV	Ac.	F_1 score	Se	Sp	PPV	NPV	Ac.	F_1 score
High activity (primiparous)	3	21	86	95.24	98.84	95.24	98.84	98.13	95.24	100	87.21	65.63	100	89.72	79.25
	4	21	82	95.24	100	100	98.8	99.03	97.56	100	89.02	70	100	91.26	82.35
	5	21	81	95.24	100	100	98.78	99.02	97.56	100	88.89	70	100	91.18	82.35
	6	21	77	90.48	100	100	97.47	97.96	95	100	89.61	72.41	100	91.84	84
Not active, Active, Eating	3	41	157	95.12	100	100	98.74	98.99	97.5	100	88.54	69.49	100	90.91	82
	4	41	150	92.68	100	100	98.04	98.43	96.2	100	89.33	71.93	100	91.62	83.67
	5	40	149	92.5	100	100	98.03	98.41	96.1	100	89.26	71.43	100	91.53	83.33
	6	40	143	90	100	100	97.28	97.81	94.74	100	90.21	74.07	100	92.35	85.1
Rumination	3	41	157	92.68	99.36	97.44	98.11	97.98	95	100	88.54	69.49	100	90.91	82
	4	41	150	92.68	100	100	98.04	98.43	96.2	100	89.33	71.93	100	91.62	83.67
	5	40	149	92.5	100	100	98.03	98.41	96.1	100	89.26	71.43	100	91.53	83.33
	6	40	143	90	100	100	97.28	97.81	94.74	100	90.21	74.07	100	92.35	85.1

Table 2. 4: Results from models performance (%) where random forest (RF) classifier was used on sensor data from an ear-attached 3-axis accelerometer (CowManager, Agis Autimatisering, Harmelen, Netherlands) device from all day using a 6 hour time window. Different cut-off values were chosen based on the highest classification probabilities.

Behavior	Time Lag *k*	Sample Size		20% cut-off						30% cut-off					
		Metritis	Non-metritis	Se	Sp	PPV	NPV	Ac.	F_1 score	Se	Sp	PPV	NPV	Ac.	F_1 score
High activity (primiparous)	7	21	82	95.24	100	100	98.8	99.03	97.56	100	89.02	70	100	91.26	82.35
	8	21	81	95.24	100	100	98.78	99.02	97.56	100	88.89	70	100	91.18	82.35
	9	21	81	95.24	100	100	98.78	99.02	97.56	100	88.89	70	100	91.18	82.35
	10	21	78	90.48	100	100	97.5	97.98	95	100	89.74	72.41	100	91.92	84
	11	21	77	90.48	100	100	97.47	97.96	95	100	89.61	72.41	100	91.84	84
Not active, Active, Eating	7	41	150	92.68	100	100	98.04	98.43	96.2	100	89.33	71.93	100	91.62	83.67
	8	40	149	92.5	100	100	98.03	98.41	96.1	100	89.26	71.43	100	91.53	83.33
	9	40	149	92.5	100	100	98.03	98.41	96.1	100	89.26	71.43	100	91.53	83.33
	10	40	145	92.5	100	100	97.97	98.38	96.1	100	89.66	72.73	100	91.89	84.21
	11	40	142	90	100	100	97.26	97.8	94.74	100	90.14	74.07	100	92.31	85.1
Rumination	7	41	150	92.68	100	100	98.04	98.43	96.2	100	89.33	71.93	100	91.62	83.67
	8	40	149	92.5	100	100	98.03	98.41	96.1	100	89.26	71.43	100	91.53	83.33
	9	40	149	92.5	100	100	98.03	98.41	96.1	100	89.26	71.43	100	91.53	83.33
	10	40	145	92.5	100	100	97.97	98.38	96.1	100	89.66	72.73	100	91.89	84.21
	11	40	142	90	100	100	97.26	97.8	94.74	100	90.14	74.07	100	92.31	85.1

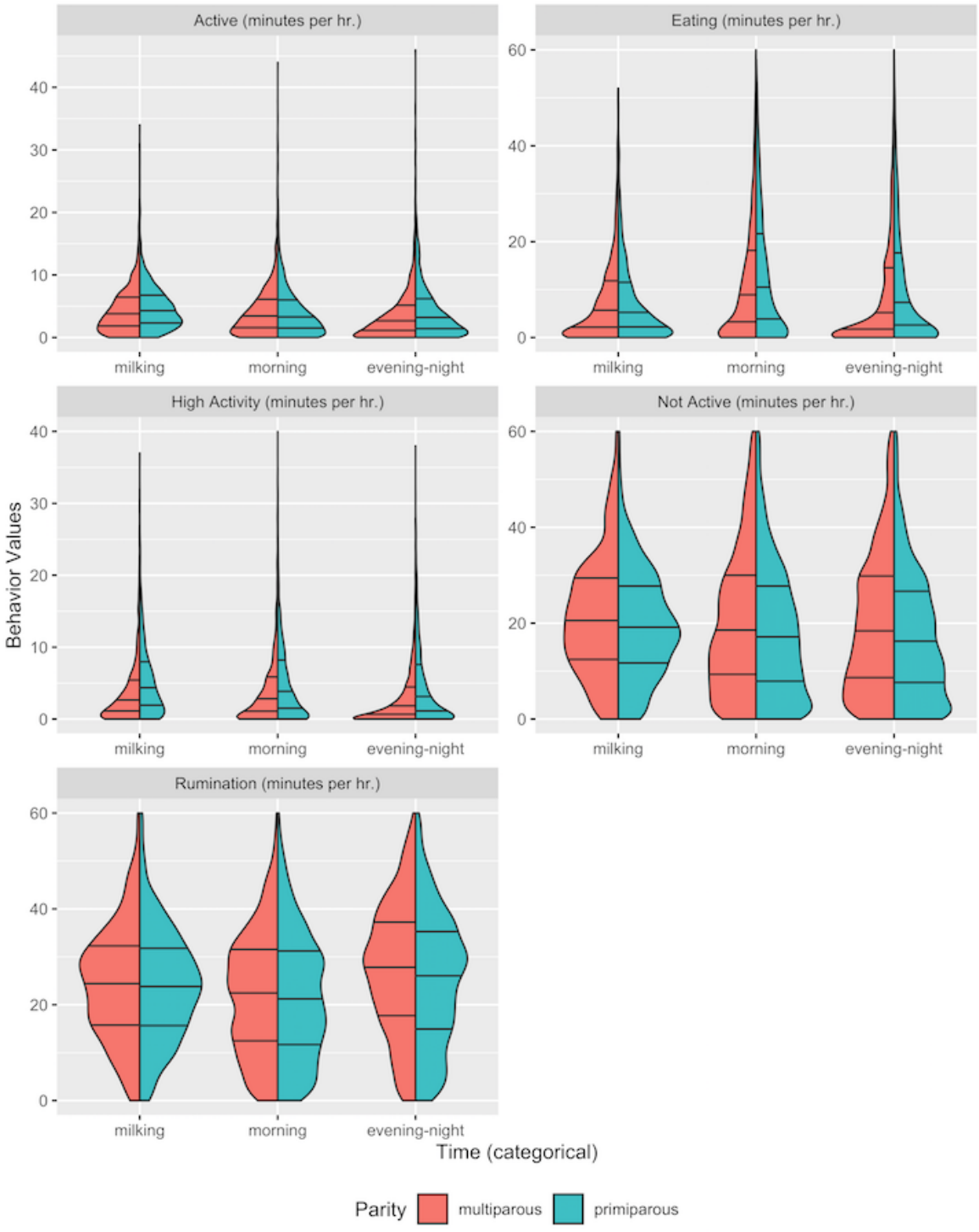

Figure 2. 1: Distribution of the raw sensor data stratified by parity and time of the day for the behavioral variables measured with an ear-tag attached 3-axis accelerometer (CowManager, Agis Autimatisering, Harmelen, Netherlands). Horizontal lines indicate mean and standard deviation. Milking is from 4:00 to 5:59 h and from 15:00 to 16:59 h; morning is from 6:00 to 14:59 h; evening-night is from 17:00 to 3:59 h in the following day.

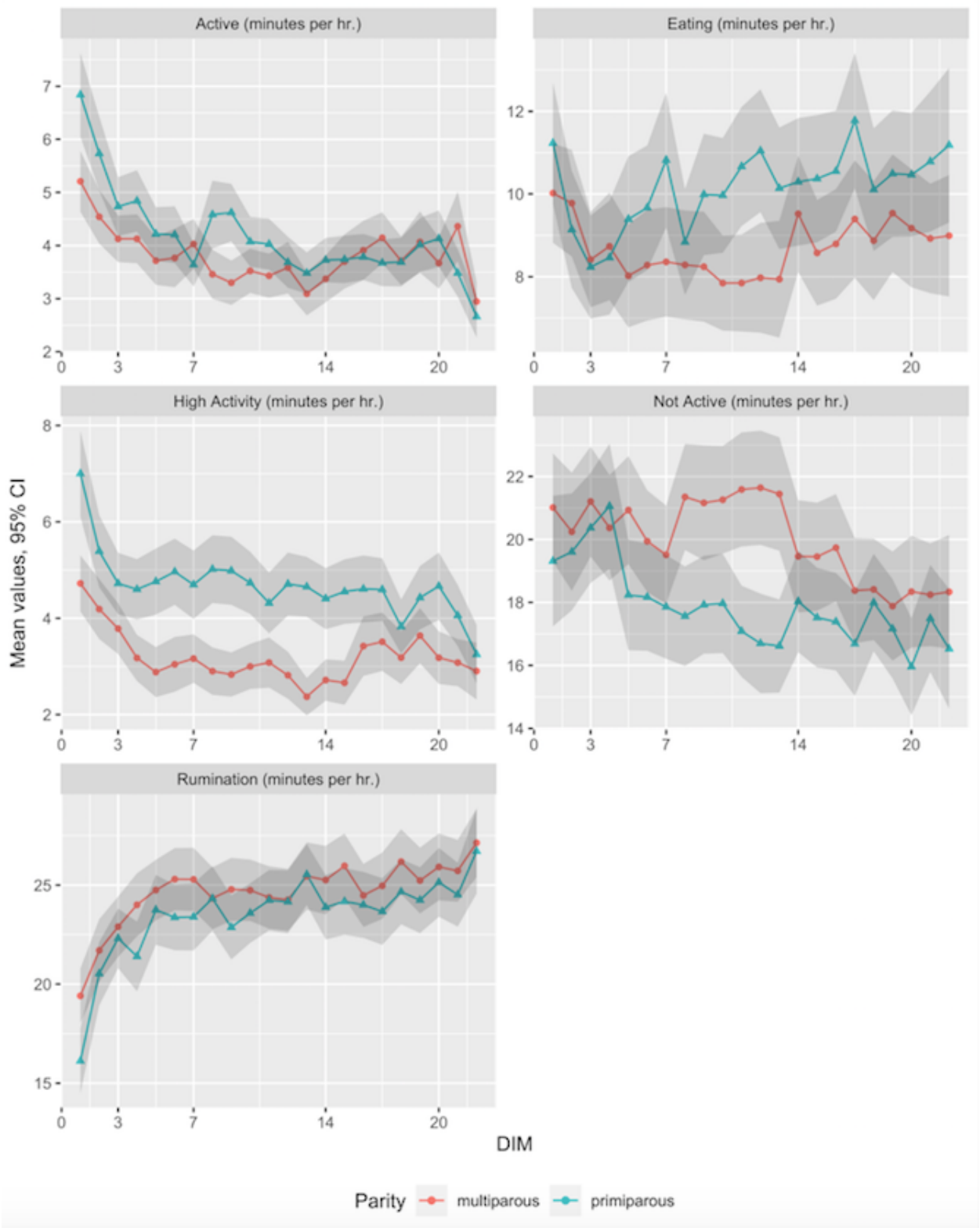

Figure 2. 2: Mean raw sensor data and 95% C.I. for the mean by days in milk (DIM) stratified by parity for the five behavioral variables registered by an ear-tag 3-axis accelerometer (CowManager, Agis Autimatisering, Harmelen, Netherlands) for the whole study period.

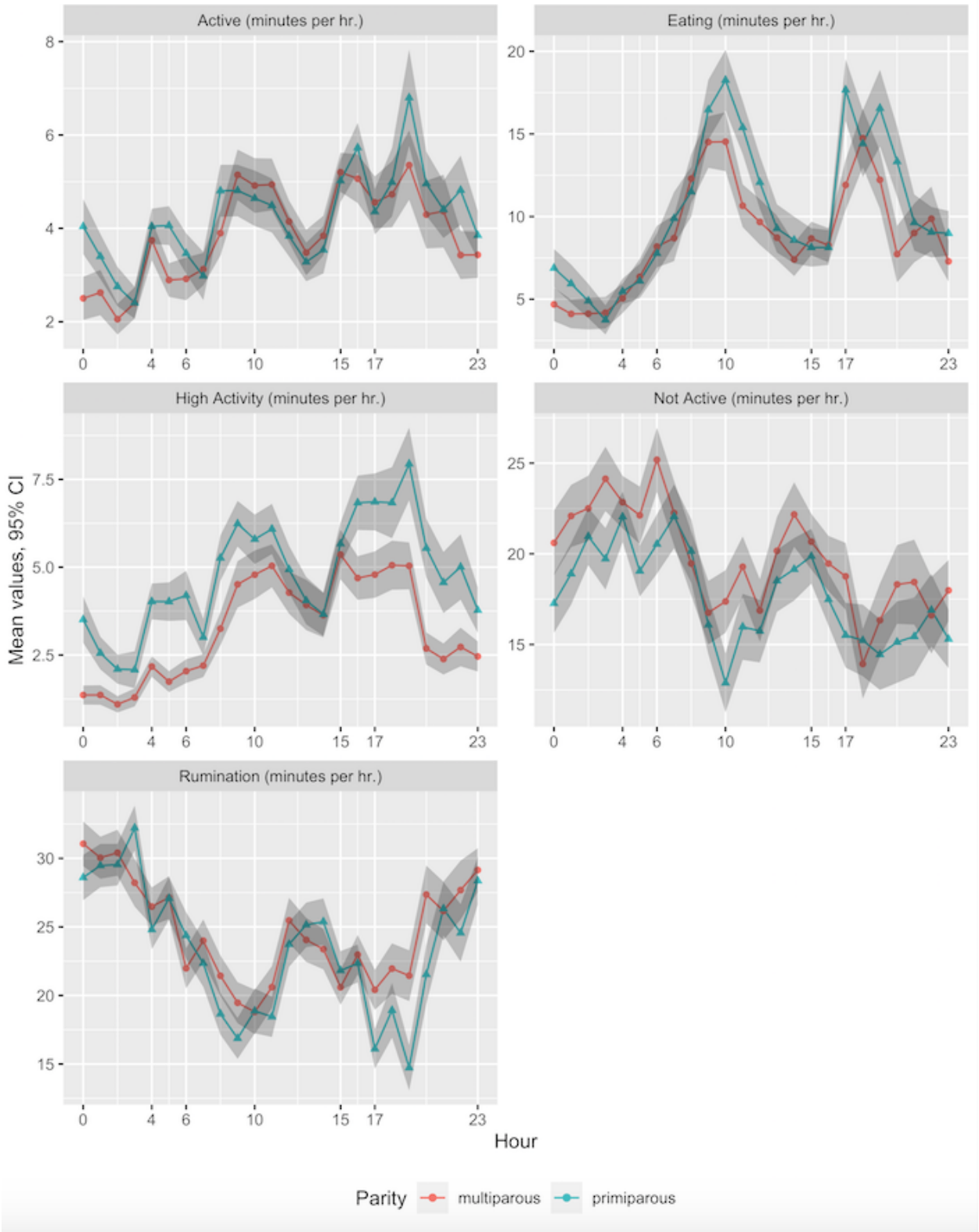

Figure 2. 3: Mean raw sensor data and 95% C.I. for the mean by the time of the day (Hour) stratified by parity for the 5 behavioral variables measured with an ear-tag 3-axis accelerometer (CowManager, Agis Autimatisering, Harmelen, Netherlands).

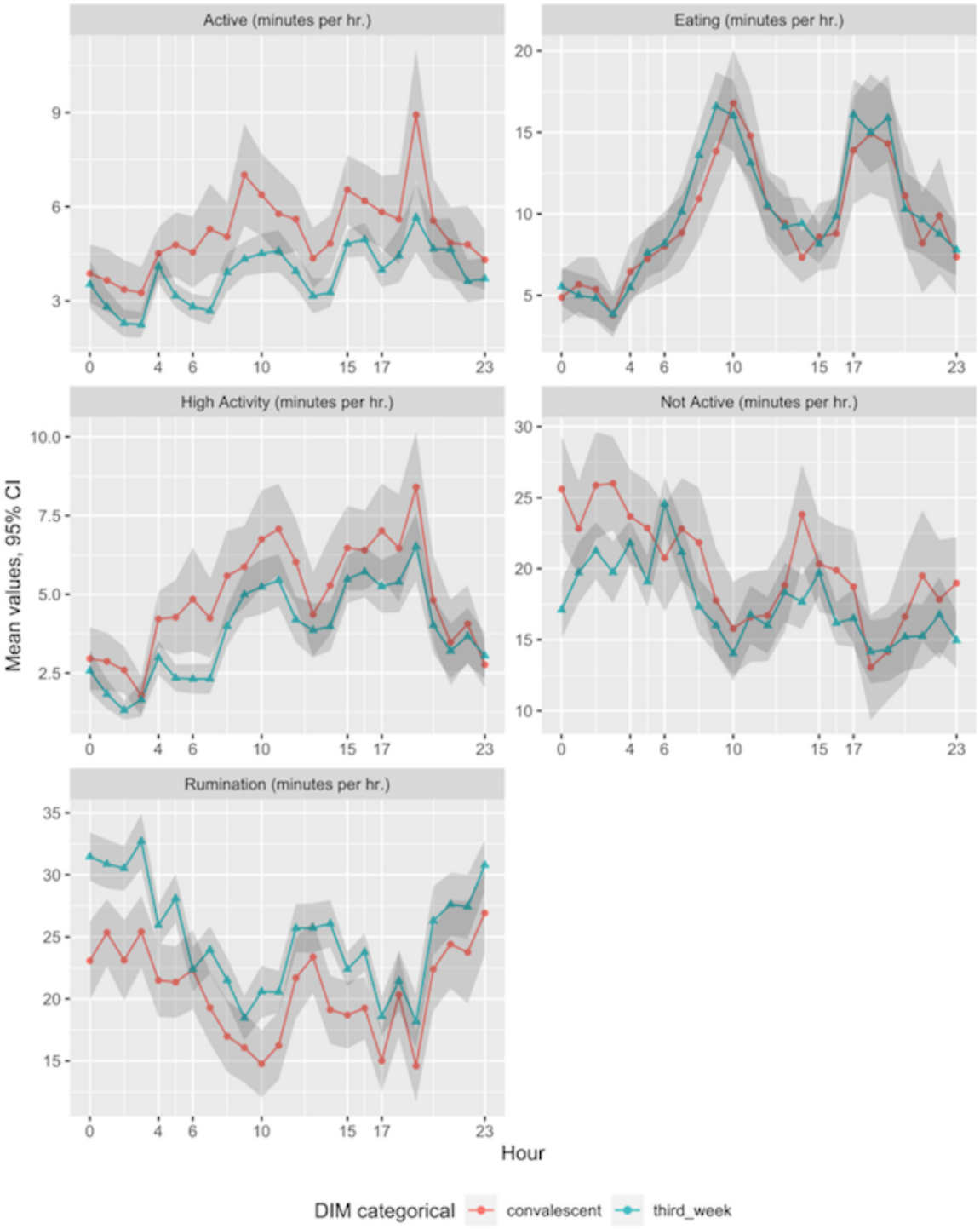

Figure 2. 4: Mean raw sensor data and 95% C.I. for the mean sensor values for each behavior measured with an ear-attached 3-axis accelerometer (CowManager, Agis Autimatisering, Harmelen, Netherlands) in a 24 hour period by parity and days in milk (DIM) categorized as convalescent (parturition to 3 DIM), first week (4 – 7 DIM), second week (8 – 14 DIM), and third week (15 – 21 DIM).

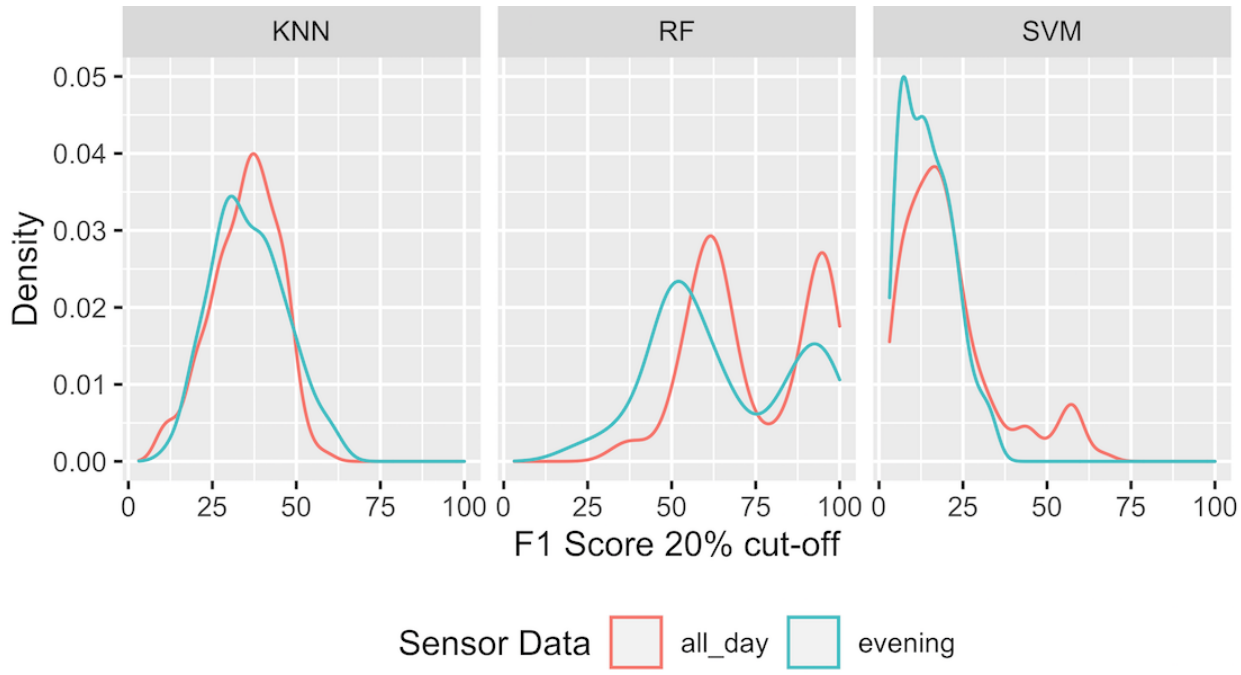

Figure 2. 5: Distribution of F_1 scores using the 20% highest class probabilities as threshold when sensor data registered by an ear-attached 3-axis accelerometer (CowManager, Agis Autimatisering, Harmelen, Netherlands) were aggregated using time windows of 24, 12, 6, and 3 hours.

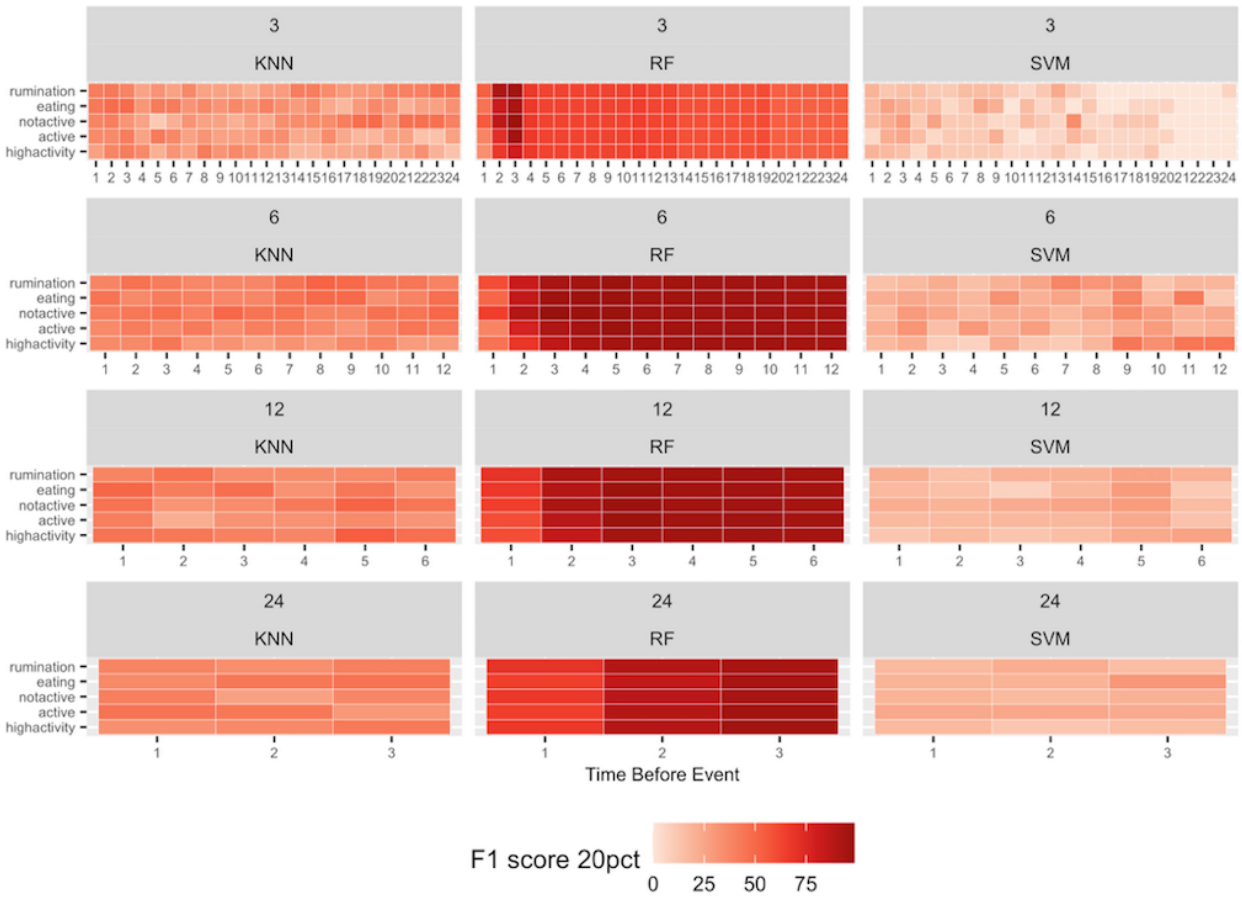

Figure 2. 6: F_1 scores using the 20% highest class probabilities as cut-off when sensor data were aggregated using time windows of 24, 12, 6, and 3 hours. F_1 scores (%) are shown for those models where all sensor data were used to fit the modes and parity was not taken into account. F_1 scores are shown for different time lags before a given metritis event for each one of the classifiers (k-nearest neighbors, random forest, and support vector machines).

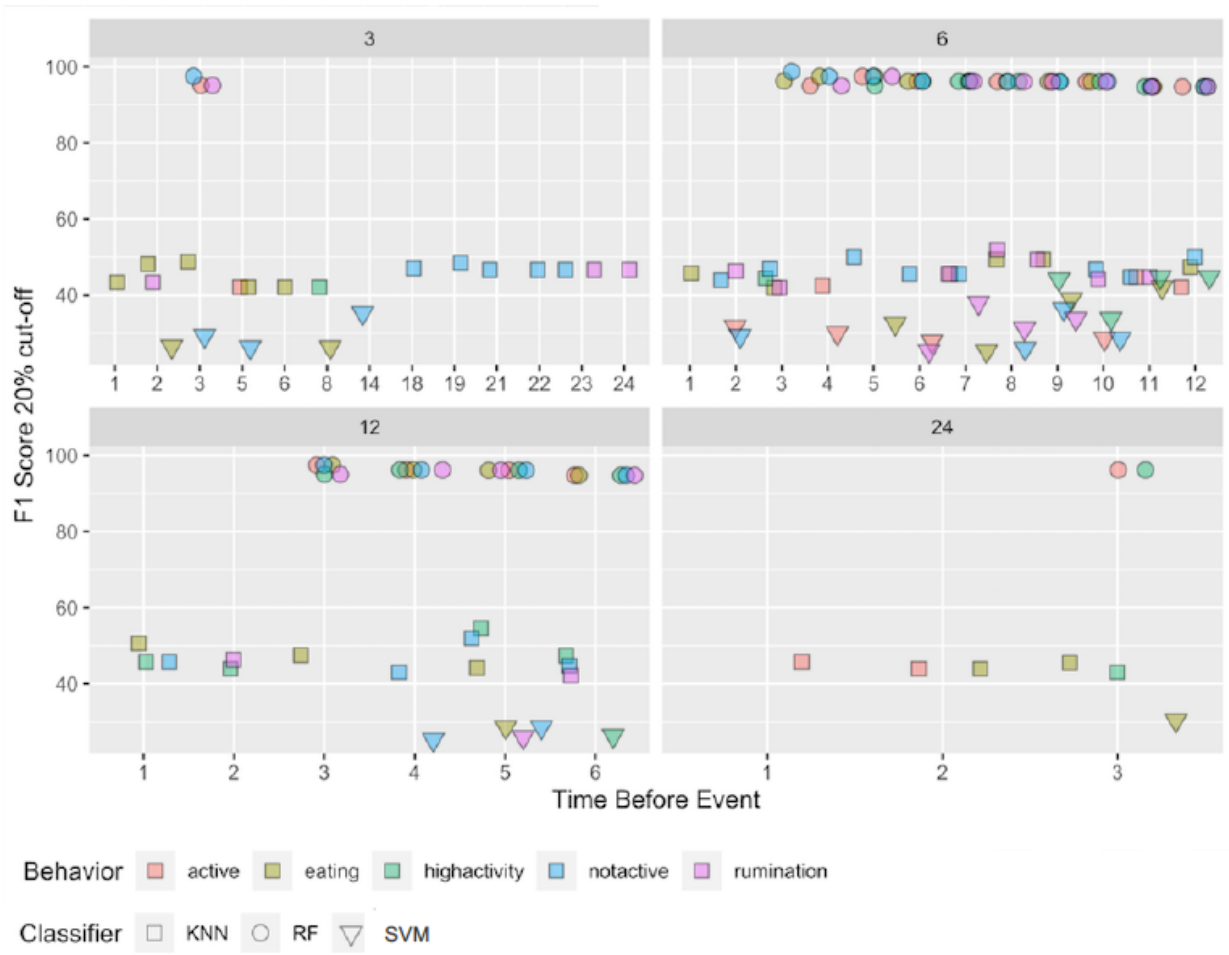

Figure 2. 7: Distribution of F_1 scores (%) using the 20% highest class probabilities as threshold from the upper quartile by behavior and classifier when sensor data registered by an ear-tag 3-axis accelerometer (CowManager, Agis Autimatisering, Harmelen, Netherlands) were aggregated using time windows of 3, 6, 12, and 24 hours, and time series sensor data from all day were used.

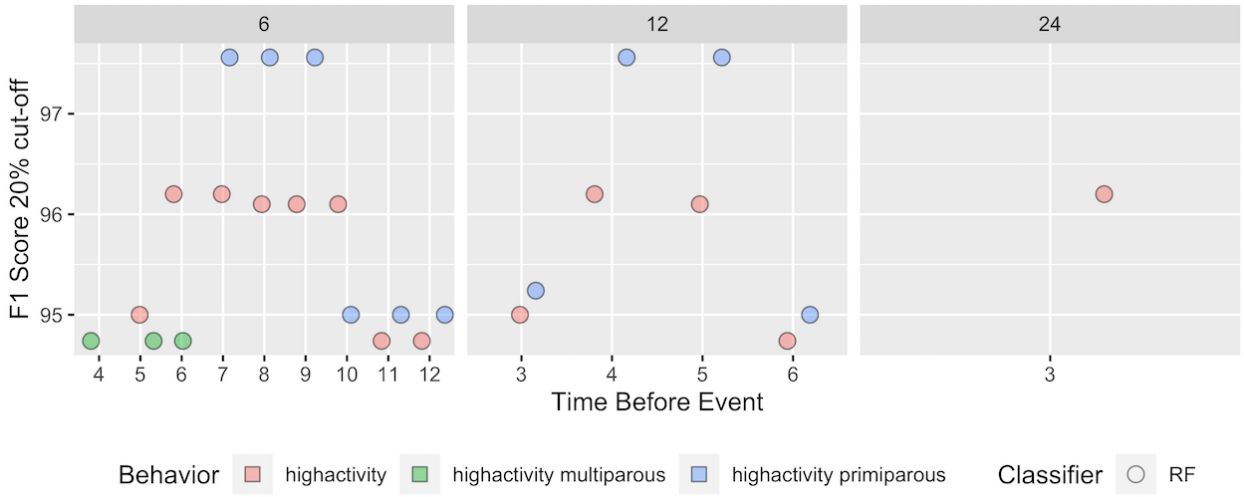

Figure 2. 8: Distribution of F_1 scores (%) for high activity behavior registered by an eat-attached 3-axis accelerometer (CowManager, Agis Autimatisering, Harmelen, Netherlands) at the 20% cut-off by classification algorithm and parity stratified by different time windows (3, 6, 12, and 24 hours).

3 Comparative performance analysis of three machine learning algorithms applied to sensor data in dairy cattle to predict metritis events: behaviors measured with a leg-attached accelerometer.

G. Vidal,[1] J. Sharpnack,[2] P. Pinedo,[3] I. C. Tsai,[4] A. R. Lee,[4] and B. Martínez-López[1]

[1] Center of Animal Disease Modeling and Surveillance (CADMS). Department of Medicine and Epidemiology. School of Veterinary Medicine. University of California, Davis, CA, 95616.

[2] Department of Statistics. University of California, Davis, CA 95616.

[3] Department of Animal Sciences. Colorado State University, Fort Collins, CO, 80523.

[4] Department of Animal Sciences. College of Agriculture, Food, and Environment. University of Kentucky, Lexington, KY, 40546.

3.1 Abstract

Routinely collected sensor data could be used in disease predictive modeling but a better understanding of its potential is needed. The objectives of this study were 1) to compare the performance of *k*-nearest neighbors, random forest, and support vector machines classification algorithms on the detection of behavior patterns associated with metritis events measured by a leg-attached accelerometer (TrackaCow, ENGS, Hampshire, UK); 2) to study whether farm scheduled activities have an impact on classifier performance; 3) to identify which behaviors yield the greatest F_1 score for metritis events prediction; and 4) to estimate the optimal level of aggregation for the hourly raw sensor data and how much behavioral data are needed in order to obtain the greatest F_1 score on metritis prediction. Data from 35 dairy cows that either did not experience any disease postpartum or were only diagnosed with metritis were retrospectively selected from a dataset containing sensor data and clinical information from 138 lactating cows during the first 21 days after parturition at University of Kentucky Coldstream Dairy from June 2014 to May 2017. A total of 188 non-metritis and 51 metritis events were created based on changes in metritis scores recorded during clinical examination. These events were associated with a total of 10,874 - 14,138 hourly sensor data from lying time, lying bouts, steps, intake, and intake visits. Sensor data corresponding to the 3 days before a metritis event were aggregated every 24, 12, 6, and 3 hours, resulting in 1,386 models. All behaviors changed throughout the study period and showed distinct daily patterns. From the three algorithms, Random Forest had the highest and most consistent performance, with no impact scheduled farm activities on classifier performance. Furthermore, 3 and 6 hours aggregation levels for the sensor data had the best balance between F_1 scores and consistency of results across different times units before

a metritis event. Based on our findings, we concluded that steps and lying time can be used to predict metritis using data from up to 2 to 3 days before a metritis event. Findings from this study will be used to develop more complex prediction models that could identify cows at higher risk of experiencing metritis, among other negative health outcomes.

Keywords: predictive modeling; classification algorithms; precision dairy farming; postpartum period; dairy cattle behavior

3.2 Introduction

Metritis is a common disease that is diagnosed in 30 to 50% of dairy cows (LeBlanc, 2010). Combined with other metabolic diseases such as hypocalcemia or hyperketonemia, post-partum infectious diseases have short- and long-term effects on welfare, reproductive health, and antibiotic use (LeBlanc, 2010). It is generally accepted that these diseases do translate into sickness behaviors as part of an adaptive response to infection or injury that help the animal to cope with the stressor. Most sickness behaviors are associated with depression, loss of appetite, and weight loss (Tizard, 2008), which can be measured with precision dairy farming (PDF) technologies such as sensor devices. Increasing herd sizes and labor cost, together with lower ratios of farm staff to animals have resulted in a rapid growth of these technologies (Rutten et al., 2013). Machine learning (ML) is a group of statistical models used in precision farming, among other fields, which goal is to find predictive patterns. Therefore, ML algorithms can be used to develop predictive models to identify which cows are at higher risk of becoming clinically ill. Given the high frequency at which changes in behavioral patterns can be analyzed when PDM technologies and ML algorithms are combined, there is potential for earlier disease diagnosis compared with traditional diagnostic methods. Resulting earlier clinical or management interventions could prevent or mitigate the impact of stress and clinical disease on animals (Weary et al., 2009; LeBlanc, 2010; Dittrich et al., 2019). Despite its potential, precision farming on metritis detection has been understudied, with only an estimated 16% of the precision farming literature being related with disease around parturition (Rutten et al., 2013).

Different authors have found reduced behaviors such as lying time (Neave et al., 2018; Sepúlveda-Varas et al., 2014; Urton et al., 2005), feeding, and rumination duration associated

with metritis during the transition period (Neave et al., 2018; Stangaferro et al., 2016; Steensels et al., 2017). However, common limitations of these studies are lack of control for concurrent postpartum diseases, behavioral data aggregation before and after disease diagnosis resulting in lost temporal relationships, and lack of consideration of within-same-day behavior variability due to farm scheduled activities (Huzzey et al., 2007b; Stoye et al., 2012). The objective of the present study was to compare the performance of three ML classification algorithms (*k*-nearest neighbors, random forest, and support vector machines) on the detection of behavioral patterns measured with a leg-attached accelerometer, associated with changes in metritis score throughout the post-partum period in dairy cows. A second goal was to identify whether farm scheduled activities had an impact on ML classification algorithm performance. A third goal was to determine which animal behaviors yield the greatest F_1 score for metritis prediction, to estimate the optimal time aggregation for the raw sensor data, and to estimate how much behavioral data are necessary to analyze for metritis prediction. Our findings would provide a base for the development of more complex prediction models. These can indicate to the farmer which cows are at higher risk of developing metritis while optimizing the use of sensor data.

3.3 Material and Methods

The data used in this study was part of a large study designed to quantify physiological and behavioral changes associated with mastitis, lameness, estrus, and postpartum diseases, using multiple PDM technologies (Tsai, 2017; Lee, 2018). The larger study included data from 138 lactating cows at the University of Kentucky Coldstream Dairy (Lexington, KY, USA) that were enrolled in the study during two different periods: the first, from June 2014 to October 2015,

and the second, from July 2016 to May 2017 under Institutional Animal Care and Use Committee #2013-119 and 2016-2368, respectively.

3.3.1 Population Data

From the original dataset, a total of 35 dairy cows that either did not experience any disease postpartum or were only affected by metritis were retrospectively selected. Cows were enrolled in the study after parturition and were followed for 21 days. Cows were excluded from the study if they died or were culled from the herd before 21 days in milk (DIM).

Details about animal management and study design are provided in a companion manuscript. Briefly, cows were moved to a close-up dry pen a month before the expected calving date, and moved again to a fresh cow pen upon parturition. Lactating cows were housed in two free-stall barns and were provided *ad libitum* access to fresh water in each barn. Lactating cows were fed the same TMR between 6:00 to 10:00 h and 12:30 to 15:00 h. The lactating diet consisted of forage, alfalfa hay, mineral and vitamin supplement, concentrate mix, whole cottonseed, and alfalfa haylage. During the second study period, feed was pushed up 22 times per day by an automated feed pusher (Lely Juno, Ley Robots, Masslius, the Netherlands). Cows were milked two times per day at 4:30 to 5:30 h and 15:30 to 16:30 h in a double 2 X 2 tandem-milking parlor.

3.3.2 Clinical Data

Disease definitions and the health-monitoring program used in the study are provided in detail in Chapter 1 (Vidal et al., Chapter 1). In short, fresh cows were monitored daily from 7:00

to 10:00 h for the first 21 days of lactation. A MetriCheck (Simero Tech Ltd, Hamilton, New Zealand) device was used to obtain a uterine discharge sample and scored on a 1 to 3 scale using a scale modified from Sheldon et al. (2006). Briefly, score 1: thick, viscous discharge, clear, opaque or red to brown in color, no odor or milk; score 2: white or yellow pus, moderate to thick discharge, milk odor; score 3: pink, red, dark red, or black watery discharge, detectable offensive odor, possibly intolerable. Cows with score $\geq$ 2 were classified as metritis cases (Tsai, 2017; Lee, 2018). Uterine discharge was scored on 3, 5, 7, 9, 11, 17, 19, and 21 DIM, and during the first study period, an additional sample was scored on 14 DIM, while during the second study period additional samples were taken on 13 and 15 DIM. Cows were also monitored for hypocalcemia, hyperketonemia, mastitis, lameness, and retained placenta. Hypocalcemia was defined as calcium level in blood serum < 8.6 mg/dL (Chapinal et al., 2011), collected by caudal venipuncture on 3, 7, 14, and 21 DIM. Hyperketonemia was defined as beta-hydroxybutyrate (BHBA) concentration in blood $\geq$ 1.2 mmol/L (Kaufman et al., 2016) measured with Precision Xtra electronic handheld device (Abbott Laboratories, Chicago, IL, USA) on days 3, 7, 14, and 21 DIM, and BHBCheck (PortaCheck Inc., Moorestown NJ, USA) on days 1, 2, 3, 4, 5, 6, 7, 10, 14, and 21 DIM for the first and second study periods, respectively. Cows were diagnosed with clinical mastitis using the following criteria: watery, thickened, and discolored milk; milk containing blood, pus, flakes, or clots; edema, erythema; or pain in the associated quarter (Royster and Wagner, 2015) between 1 and 21 DIM by trained milkers. Furthermore, subclinical mastitis was assessed measuring somatic cell count (SCC) on days 4 $\pm$ 2 DIM and 9 $\pm$ 2 DIM via flow cytometry in quarter milk samples. Cows with SCC $\geq$ 200,000 cells/mL in one or more quarters were considered positive for subclinical mastitis. Finally, locomotion scores were recorded on days 1,

7, 14, and 21 postpartum on a 1 to 3 scale (Schlageter-Tello et al., 2014). Retained placenta was recorded if fetal membranes were retained for > 24 hours (Sheldon et al., 2006).

For any given cow and day, a metritis event was assigned when a cow was getting or being with metritis, this is, the metritis score increased, changed from 3 to 2, or when the score remained 2 or 3, between two consecutive uterine discharge evaluations. Similarly, for any given cow and any given day, a non-metritis event was assigned when a cow recovered from metritis or stayed healthy, this is, when the metritis score changed or remained as 1, between two consecutive uterine discharge evaluations. Diagnosis of metritis was assigned to happen at 6:00 h on each one of the days when uterine discharge was evaluated for time series data manipulation purposes. To study the effect of scheduled farm activities, models were also fitted using only sensor data from 17:00 h to 3:00 h (evening-night models) and therefore, diagnosis of metritis was assigned to happen at 17:00 h on each one of the days when uterine discharge was evaluated for later time series data manipulation.

3.3.3 Sensor Data and Data Pre-processing

For this study, information per cow included five different behaviors measured from parturition to 21 days postpartum with a leg-attached 3-axis accelerometer (TrackaCow, ENGS, Hampshire, UK) that records hourly data regarding lying time (minutes per hour), lying bouts (number per hour), steps (number per hour), intake (minutes per hour), and intake visit (number of visits to the feedbunk per hour). This device has been previously validated by Borchers (2015), Borchers et al. (2016), and Chapinal et al. (2007).

Time series sensor data consisted on the hourly measurements for each behavior i corresponding to the 3 days prior to each metritis event, assigning the time of diagnosis t at 6:00 h on each one of the days when uterine discharge was evaluated. Therefore, the 6:00 h time was used as offset for later transformations of the time series sensor data. When only sensor measurements corresponding to evening-night were used, for any given day, only sensor data from 17:00 to 3:00 h were considered, being the time of diagnosis t assigned at 17:00 h on each one of the days when uterine discharge was evaluated. The following time series data transformations were applied to both time series: one with observations for every hour, and another one containing only those corresponding to the evening-night hours.

The first time series data transformation was to remove seasonality by differencing the time series. In order to do that, we subtracted for each cow, behavior i, and hour within a 3 day period before a given metritis event, the measurement registered by the sensor in the previous 24 h from each hourly sensor measurement. The time series data for each metritis event at time t was defined by:

$$\left(x_{i,t-1}, x_{i,t-2}, \dots, x_{i,t-n}\right)$$

where:

x_i was the differenced hourly sensor measurement for behavior i and time t, being $i \in \{lying, lying\ bouts, steps, intake, intake\ visit\}$

n was the time step within a 3 day (or 72 hours) period.

Next, we transformed the time series sensor data by aggregating the differenced hourly measurements using the mean of the time window tw_1. In order to assess classifier performance at different levels of sensor data aggregation, we used 4 different widths to compute the mean:

3, 6, 12, and 24 h. As result, the new time series data for each metritis event at time t was defined by:

$$\left(\bar{x}_{ij,t-1}, \bar{x}_{ij,t-2}, \ldots, \bar{x}_{ij,t-m}\right)$$

where:

$\bar{x}_{ij}$ was the mean sensor value for behavior i and time window tw_1 of width j,

being $i \in \{lying, lying\ bouts, steps, intake, intake\ visit\}$,

and $j \in \{3\ h, 6\ h, 12\ h, 24\ h\}$

m was the time step within a 3 day period. The number of time steps that could be included within a 3 day period was a function of the width j of the time window tw_1.

3.3.4 Model Fitting

We selected the number of model inputs (or features) by using a time window tw_2 of width k. In order to assess classifier performance at different widths, we used multiple values for k within a 3 day period before each metritis event. Therefore, the model inputs for each model were:

$$\left(\bar{x}_{ij,t-1}, \bar{x}_{ij,t-2}, \ldots, \bar{x}_{ij,t-k}\right)$$

where the width $k = 1, 2, \ldots, l$, and l was the number of time steps included as features within a 3 day period before a given metritis event.

The number of features in our models ranged from 1, when sensor data were aggregated with a tw_1 width j of 24 hours and tw_2 width k of 1, to 24 features when sensor data were aggregated using a tw_1 width j of 3 hours and tw_2 width k of 24, corresponding to 72 hours prior to the event.

In this paper, we evaluate the ability of 3 supervised ML classifiers (*k*-nearest neighbors, random forest, and support vector machines) to discriminate among 2 possible distinct patterns (metritis and non-metritis events) in 5 animal behaviors. Within each one of the 5 behaviors, for each one of the three classifiers, and for each combination of sensor data aggregation and number of time units before the metritis event, different models were fitted. Details about each classifier are provided in a companion manuscript. Briefly, *k*-nearest neighbors (*k*-NN) relies on the assumption that similar data points exist in close proximity and estimates the closeness using Euclidean distance for each data point to the rest of the data points (Fix and Hodges, 1951; Dasarathy, 1991; Hastie et al., 2009). Random forest (RF) is made up from many decision trees, a flowchart of questions asked about the data that leads to a predicted class (metritis or non-metritis events) with the greatest reduction in Gini Impurity, or the probability that a randomly chosen sample in a set of data points or node would be correctly labeled if it was labeled by the distribution of samples in the node (Breiman, 2001; Hastie et al., 2009). In contrast, support vector machines (SVM) estimates the optimal hyperplane, or decision hyperplane, that separates the different classes while maximizing the distance, or margin, to the closest point from either class, also called support vectors. One of the advantages of SVM is the use of the kernel function, a mathematical function that transforms the feature space to deal with cases in which classes are not linearly separated (Vapnick, 1995; Hastie et al., 2009).

For each one of the three ML classifiers, one model was fitted for each combination of behavior l, time window tw_1 of width j, and time window tw_2 of width k.

3.3.5 Model Performance

To assess model performance, we used fivefold cross-validation (5-FCV). Specifically, for any given model, 4/5ths of the data were used to fit the model, whereas the other 1/5th was used to calculate the prediction error of the fitted model. This process was repeated each time until all 5 folds had been used for both, fitting the model and validation, resulting in an average prediction error. Grid Search (GS) was used as a strategy to optimize the classifier, except for RF classifier, where GS was performed after Randomized Search (RS) in order to reduce the grid search so computing time was manageable. Optimal parameters that were found to allow for best mean cross-validation accuracy were used to train the final model (Table 3.1). After optimization, the prediction class probability for each health event of being classified as metritis was obtained and ranked from highest to lowest, and the top 20, 30, and 40% class probabilities were used as different cut-off points. For each cut-off point, classification performance was evaluated using estimates of sensitivity (Se or recall), specificity (Sp), positive predictive value (PPV or precision), negative predictive value (NPV), accuracy (Ac), F_1 score, the area under the curve (AUC) for the receiver operating characteristic (ROC) curve and Precision Recall (PR)-curves. Sensitivity is estimated as the ratio of correctly predicted positive observations to all observations in the actual class (metritis event). Specificity is estimated as the ratio of correctly predicted negative observations to all observations in the actual class (non-metritis event). Positive predictive value is the ratio of correctly predicted positive observations to all predicted positive observations. Similarly, NPV is the ratio of correctly predicted negative observations to all predicted negative observations. Accuracy is the ratio of correct predictions to all number of observations. F_1 score is the weighted average of PPV and Se. This score takes both false

positives and false negatives into account, and it is more useful than Ac in situations where the distribution of the observations in each class is unbalanced. F_1 score was computed as $(1 + b^2)*(PPV * Se)/((b^2 * PPV) + Se)$, where b = 1 (Saito and Rehmsmeier, 2015).

Classifier implementations were taken from the open-source Python library scikit-learn (Pedregosa et al., 2011). The feature extraction and the optimization of the classifier parameters were implemented using Python programming language, version 2.7 (Python Software Foundation, http://www.python.org). Plots were done using ggplot2 library (Wickham, 2009), using R open-source statistical software (R Core Team, 2017).

3.4 Results

A total of 35 dairy cows (Jersey = 20; Holstein = 15; primiparous = 17; multiparous = 18) were retrospectively selected from the original dataset (n = 138) containing clinical and sensor data from parturition to 21 DIM. Average $\pm$ SD milk yield was 36.1 kg. $\pm$ 15.6. Of the 35 cows selected, 13 did not have any metritis events during the study period, while 22 were diagnosed at least once with metritis (score $\geq$2), occurring at 12 DIM (12.02 $\pm$ 4.72 DIM). Among these, 2 cows had retained fetal membranes and were kept for data analysis. None of the selected animals had hyperketonemia, mastitis, or hypocalcemia. The proportion of metritis events for primiparous and multiparous were 20% and 23%, respectively. Based on the changes of metritis score between two consecutive evaluations, 239 health events were created, and of those, 188 were in the non-metritis event class, while 51 were in the metritis event class, resulting in an unbalanced dataset. The number of hourly sensor records ranged from 10,874 for behaviors intake and intake visit, to 14,138 for lying, lying bouts, and steps.

Cows showed high variability in their behaviors during the study period, especially regarding number of steps (98.8 ± 72.51 number/h) and lying time (21.74 ± 21.06 min/h), followed by intake (7.54 ± 12 min/h). This trend was constant regardless of the level of sensor data aggregation and time of the day (Table 3.2). Furthermore, the distributions for lying bouts, steps, intake, and intake visits, were right-skewed, and differences in the mean values by parity were greater during the evening-night hours for lying and steps (Figure 3.1).

3.4.1 Changes in Behavior by Days in Milk and Time of Day

There were changes across all behaviors from parturition to 21 DIM, with significant changes in the first 3 days post-partum for some of the behaviors. During the first 3 DIM, lying time increased while lying bouts and steps decreased. Overall, intake showed an upward trend throughout the study period. When behaviors were stratified by parity, multiparous cows showed significantly lower number of lying bouts than primiparous cows throughout the study period, while significant differences in number of steps by parity occurred around 7 and 14 DIM. During the whole study period, primiparous cows tended to spend less time lying down with greater number of steps than multiparous cows (Figure 3.2). When looking at the variability of each behavior throughout the day for the whole study period, lying time and steps had inverse trends, with greater number of steps during milking times and at 10:00 h, time at which lying bouts were also the greatest. Intake and intake visits showed similar trends, with greater values right after milking times. Differences by parity were observed for lying bouts throughout the day, while differences by parity regarding lying and steps were observed during the evening-night

hours, when multiparous spent more time lying down and took fewer steps than primiparous (Figure 3.3).

3.4.2 Changes in Behavior by Time of Day Stratified by Days in Milk

To further explore the changes of the different behaviors across the study period, we also looked into the variation throughout the day stratified by DIM. We categorized DIM into 3 distinct periods: convalescent (from parturition to 3 DIM), first week (4 to 7 DIM), second week (8 to 14 DIM), and third week (15 to 21 DIM). No significant differences were observed across the different periods, however, intake and lying tended to be greater while steps tended to be lower during the third week compared with the convalescent period. It is worth noticing that such trends became unnoticeable during milking times and, in some cases, around 10:00 h (Figure 3.4).

3.4.3 Classifier Performance

The total number of models fitted is described in a companion paper. Briefly, a total of 1,386 models were fitted to account for differences by parity (primiparous, multiparous), by level of sensor data aggregation defined by the width j of the time window tw_1 (3, 6, 12, and 24 hours), by number of time steps before a metritis events included as features within a 3 day period (width k of the time window tw_2), and by time of the day (all day and evening-night only time series data). Of these, 45 models were fitted for each combination of behavior and classifier using sensor data regardless of the time of the day (all day), resulting in a total of 675 models. From the final number of models, 21 were fitted for each combination of behavior i and

classifier when sensor data pertaining to evening-night hours were used, resulting in a total of 315 models. Models by parity were also fitted for lying (evening-night only time series data), lying bouts (all day and evening-night only time series data), and steps (evening-night only time series data) behaviors. Fivefold cross-validation F_1 scores at the 20% cut-off were used to compare across different classifiers. For all classifiers, higher F_1 scores were obtained when sensor data were used regardless of the time of the day. Figure 3.5 shows an overview of the distribution of F_1 scores at the 20% cut-off at different times before the health event, stratified by the different levels of sensor data aggregations. Random forest had the greatest and most consistent F_1 scores across multiple levels of time aggregation and time before an event, followed by *k*-NN and SVM. Metrics performance from all models can be found in a data repository (Vidal et al.).

To better understand the performance of each classifier, we looked further into the models in the upper quartile of the F_1 score distribution at the 20% cut-off, when sensor data for all day were used. For RF, the upper quartile for F_1 score values at the 20% cut-off were between 89.93% and 97.67%, while the upper quartile for *k*-NN was between 45.39% and 60%. In contrast, the top 25% values for SVM F_1 scores were between 22.93% and 57.89% (Figure 3.6). Figure 3.7 shows the best models considered at the different levels of time aggregation and number of time steps included in the model before a health event. Our results confirmed that, among the three classifiers, RF had the best performance and, for each classifier, a greater number of behaviors with slightly greater F_1 score values ranked in the upper quartile as the level of sensor data aggregation became smaller. When data were aggregated using 24 hour time windows, the predominant behaviors were steps and intake visit when using *k*-NN or SVM

classifiers. When sensor data were aggregated using 12 hour time windows, the predominant behaviors were steps, followed by intake, lying, and intake visit. Again, the majority of models were classified using *k*-NN and SVM. When sensor data were aggregated using 6 or 3 hour time windows, lying and steps were the predominant behaviors when RF was used. In contrast, for the same level of time aggregation, intake and intake visit were the predominant behaviors when *k*-NN or SVM were used. Among the behaviors for which separate models were fitted by parity, only lying bouts ranked in the upper quartile of the distribution of F_1 score at the 20% cut-off, regardless of classifier. Our results showed that greater F_1 scores were obtained for lying bouts in primiparous compared with multiparous (Figure 3.8).

3.4.4 Best Classifier, Time Window, and Time Lag

In our study, the best balance between high F_1 score values, number of behaviors ranking amongst the best models, and consistency regarding the number of time steps before an event included in the model a given behavior ranked amongst the best models was found when RF was used with sensor data aggregated using 6 or 3 hour time windows. For the 6 hour time window, best models were found between 30 to 72 hours before the event (number of time steps before event from 5 to 12). Similarly, for the 3 hour time window, best models were found between 18 to 72 hours before the event (number of time steps before event from 6 to 24). Table 3.3 and 3.4 show the performance metrics for the selected best models at two different cut-off points. For the selected times, Se and PPV decreased as we increased the number of time steps before the health event, with their greatest values at 36 and 18 hours and 6 or 3 hour time window for the sensor data aggregation, respectively. Using the estimated predicted probabilities, we

compared the metritis events identified by RF with the clinical data. We found that the number of missed events ranged between 1 and 4, increasing as we increased the number of time steps before the health event, and none of them were two consecutive missed events, this is, the metritis had either been diagnosed before, or it was diagnosed at the following metritis evaluation.

3.5 Discussion

In this paper, we compared the performance of three different classification ML algorithms on five different behavior variables. Our goal was to assess the ability of the classification algorithms to identify patterns in the sensor data that may be associated with changes in metritis score during the first 21 days post-partum. To preserve the time structure of the sensor data, metritis events were created and sensor data from the 3 days prior to an event were aggregated at different time windows. Furthermore, to deal with the challenge of the unbalanced dataset, we used the F_1 score to evaluate classifier performance based on the predicted classification probabilities ranked from high to low.

Based on our results, behavior data can be highly variable. From the results summarized in Table 3.1 stands out that cows spent, on average, 8.7 hours/day lying down (21.74 $\pm$ 21.06 min/h), had 14.88 lying bouts per day (0.62 $\pm$ 0.85 per h), and took 2,371.2 steps per day (98.80 $\pm$ 72.51 per h). Our findings are similar to those found by others, although mean lying time was found to be in the lower of what is recommended (Bewley et al., 2010; Gomez and Cook, 2010). Differences in the mean values across studies could be due to differences in the devices used or the average DIM of the animals. Most of the studies that report descriptive statistics of different

behaviors are validation studies where cows across the whole lactation were used, increasing the average DIM of the animals in the study. This is particularly relevant since cow's behavior is constantly changing postpartum. Furthermore, differences in management practices such as high frequency feed delivery will translate into differences in lying time and lying bouts across studies (Mattachini et al., 2019).

In this study, animal behavior changed according to DIM. During the first 3 DIM, lying bouts and steps behaviors had a downward trend while lying time had an upward trend. Overall, intake had an upward trend for the whole study period. We also found that multiparous had a lower number of lying bouts and steps than primiparous, while the amount of time lying was greater than that found in primiparous, particularly during evening-night hours. The trends observed during the study period are in agreement with those found by other authors. Lying time decreased in the days following parturition, with increasing lying time as DIM increased (Chaplin and Munksgaard, 2001; Bewley et al., 2010). Udder discomfort or high demand for food have been proposed as explanations for this trend (Chaplin and Munksgaard, 2001). Feeding behavior has been found to decrease by 35% over the 2 weeks before calving and to increase by 99% over the 3 weeks following parturition (Urton et al., 2005). Differences by parity regarding lying bouts, lying time, and number of steps have been found in other studies, where primiparous cows have shown increased lying times among grazing dairy cows (Sepúlveda-Varas et al., 2014), as well as in free-stall housed cows (Vasseur et al., 2012; Barragan et al., 2018; Neave et al., 2018). In contrast, multiparous cows had greater lying times in our study, a finding supported by Piñeiro et al. (2019). It is not clear why different studies yield contradictory results for the interaction between parity and lying time, but it is possible that different findings may be

attributed to inflammatory response differences by parity (Humblet et al., 2006; Piñeiro et al., 2019), or to social dominance dynamics between primiparous and older cows (Sepúlveda-Varas et al., 2014). Nevertheless, we found that classifier performance for lying time by parity was not superior to that one in which data from all cows were pooled together.

The studied behaviors also changed according to the time of the day. When behaviors were observed in a 24 hour period, cows showed inverse trends regarding lying and steps. Lying time is a resting state that was higher during night hours, followed by the hours between morning and afternoon milking. In contrast, steps is an activity state that was higher during milking times and at 10:00 h, time at which cows were being moved to be treated or checked. These trends are supported by circadian cycle research (Ruckebusch, 1972) and similar findings have also been reported by other authors, although small differences can be found across studies due to differences in milking times, feeding management, or environmental temperature (Overton et al., 2002; DeVries and Von Keyserlingk, 2005). Differences by parity were only observed during the evening-night hours, a fact that could support the hypothesis that when left alone, cows may have greater opportunities to express their natural behavior and, therefore, using sensor data from evening-night only hours would increase the performance of classifiers. Our findings regarding model performance comparing sensor data regardless of the time of the day *vs.* data from evening-night hours only did not support this hypothesis. Nevertheless, future studies should evaluate classifier performance under different scenarios on a case basis, as there are some behaviors that may not be worth considering given certain times of the day such as milking times, where animals will not lay down or eat.

To better understand the dynamics of cow behavior throughout the study period, we looked at the behaviors in a 24 hour period when DIM was categorized. Based on our findings, we did not find significant differences across the weeks of the study. This is opposite to what we found in our companion paper, where differences were found between the convalescent period and the third week. However, based on our results, we proposed that the inclusion or exclusion of data from the first 3 DIM should be routinely evaluated in these types of studies, since results may change depending on the type of sensor device used and the nature of behavioral data being collected.

One of our objectives was to compare the performance of three different ML classifiers. Our approach to model performance evaluation with unbalanced datasets is to use F_1 score *vs.* Ac or ROC curve, two metrics that are commonly used in precision dairy farming but are not appropriate when datasets are unbalanced. When non-disease events are the majority class, Sp is expected to be high regardless of how good or bad the classifier is with the type of data at hand. Another advantage of the F_1 score is that, in order to account for the different costs different misclassification errors may have, weights for PPV and Se can be modified using the formula $F_b = (1 + b^2) * (PPV * Se) / (b^2 * PPV) + Se$. In this study, we also explored different classification cut-off probabilities to account for the trade-off between Se, Sp, PPV, and NPV. Interested readers will find all the estimated metrics at different cut-offs for each one of the different model specifications in a data repository (Vidal et al.). Based on the F_1 score distribution and consistency of results at the 20% cut-off, RF had the best performance, followed by *k*-NN and SVM, with slightly greater F_1 scores as the level of time aggregation became smaller (e.g. 3 hour time window), a finding also reported in other studies (Martiskainen et al., 2009). In

this study, amongst those models with best performance, *k*-NN achieved an F_1 score with values between 45.39 – 60%, while SVM yielded an F_1 score between 22.93 – 57.89%. In contrast, the best RF models had F_1 scores in the range between 89.93 – 97.67%. Random forest is based on decision trees, a classification method that has been used in the precision dairy farming with great success to study grazing cattle behavior (Williams et al., 2016), to predict fertility and improve heat detection in dairy cows (Caraviello et al., 2006; Vanrell et al., 2014), to predict mastitis (Kamphuis et al., 2010a), or to understand complex relationships between metabolic diseases postpartum and culling risk (Probo et al., 2018). Random forest can handle large data sets with a high number of features; however, the decision trees the RF is made of are not intuitive, making it harder to grasp the relationship existing in the input data when compared with other methods.

In dairy cattle, increased physical activity is a sign of estrus (Firk et al., 2002) and a sign of sickness behavior when decreased before and beyond metritis diagnosis (Liboreiro et al., 2015; Stangaferro et al., 2016a; Steensels et al., 2017). Measured with accelerometers that transform acceleration into angles, when attached to the leg, changes in angles are interpreted either as steps or lying. In our study, number of steps had a Se that ranged between 86.36 to 93.33%, PPV between 93.02 and 100%, and Ac between 91.57 and 98.54%, being these estimates similar when sensor data were aggregated every 6 or 3 hours. These performance metrics were greater than those reported by Mayo et al. (2019) for heat detection, although their sample size was smaller and they did not use a ranked-based approach to evaluate model performance. Our findings were also higher than those reported by Stangaferro et al. (2016), with average Se of 53% and a maximum of 70% Se for those cows with rectal temperature $\geq$ 40.0 ºC. However,

comparison is not straightforward since performance metrics provided by other authors were for the associations between metritis diagnosis and a health index, computed with proprietary algorithms that combined rumination and activity measured in arbitrary units per day. Furthermore, no values for PPV were reported since no specific disease was provided in the alert generated by their device.

Among behaviors considered as resting state, lying time has a critical role in the production potential and welfare status of dairy cattle. Associated with disease, increased lying time has been found in animals with metritis as a consequence of depression (Barragan et al., 2018), while it has been found to decrease associated with mastitis due to discomfort while lying down (Siivonen et al., 2011). In this study, lying time Se, Sp, PPV, NPV, and Ac were 90.48 – 95.45%, 98.71 – 100%, 94.87 – 100%, 97.55 – 98.82%, and 97.04 – 99.06%, respectively, with slightly greater values when sensor data were aggregated using a time window of 3 hours. Our performance metrics are higher than those found in accelerometer device validation studies, with Se, PPV, and Ac of 80%, 83%, and 84%, respectively (Martiskainen et al., 2009), as well as higher than those found using lying time 1 week before calving to predict metritis post-partum, with reported Se and Sp of 75% and 66.67%, respectively (Patbandha et al., 2012).

The number of models that ranked amongst the best ones changed based on the different levels of time aggregations and classifier. We also found that, even though intake and intake visit did not yield high F_1 scores, SVM and *k*-NN classifiers performed better with behaviors intake and intake visit while RF performed better with behaviors lying and steps. This supports the idea that some ML classifiers may work better than others for certain behaviors, and alternative ML algorithms for feeding related behaviors measured with Trackacow device

should be explored. Based on our findings, best results were obtained with sensor data aggregated using 6 or 3 hour time windows, being the 6 hour time window better for steps, while the 3 hour time window resulted in better performance for lying bouts. For optimal performance, sensor data from the previous 30 – 72 hours before the event were needed when sensor data using 6 hour time windows, although when data were aggregated every 3 hours, data from the previous 18 hours before an event did suffice. This is in agreement with what has been found by other authors: steps have been found to change 2 days before diagnosis of metritis (Steensels et al., 2017), metabolic, or digestive problems (Edwards and Tozer, 2004). Similarly, lying bouts have been found to change 2 to 3 days before metritis diagnosis (Neave et al., 2018; Piñeiro et al., 2019). Nevertheless, the appropriate combination of number of observations used as cut-off, level of sensor data aggregation, and number of time steps before metritis event to be included as features should be chosen on a farm case basis, and should be dynamically adjusted to reflect changes in the incidence of metritis cases, costs for medical treatments, and cost of missed metritis cases.

Limitations remain with current prediction models regarding how to deal with cases where multiple illnesses are present, understanding how classifier performance can change by adding multiple behaviors or devices at the same time, and how model performance translates into commercial farms. Further studies are also needed in order to identify other ML methods that have optimal performance when feeding behaviors measured with Trackacow are used.

3.6 Conclusions

The findings of this study have a number of practical implications. Our results indicate that rank-based methods for model fitting yields superior results to those studies where data were artificially balanced. Therefore, rank-based methods should be preferred when developing predictive models that will be implemented in the future. We also found that data from the last two days regarding steps and lying time measured with Trackacow device could be used to predict metritis events with RF classifier when sensor data were aggregated using either 6 or 3 hour time windows.

3.7 Acknowledgements

The authors would like to thank the University of Kentucky Coldstream dairy staff, and to all the students who helped with the fresh cow exam and data collection. We would also like to thank Jeffrey Bewley for facilitating data sharing. The work was partially supported by NSF award IIS-BigData-AI-1838207. JS is partially supported by NSF DMS 1712996.

Table 3. 1: Hyperparameter values used for optimization of k-nearest neighbors (k-NN), random forest (RF), and support vector machines (SVM) classification algorithms used on behavioral variables measured with a leg attached 3-axis accelerometer (TrackaCow, ENGS, Hampshire, UK).

Classifier	Hyperparameter	Randomized Search	Grid Search	Models Used[1]	Optimum Value
k-NN	*k*[2]	N/A	1 to 15	tw_1 j = 24h, tw_2 k = 1, 2, 3	7
RF	Bootstrap[3]	True, False	True	tw_1 j = 24h, tw_2 k = 2	True
	Max. depth[4]	10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, None	5, 10, 15, 20	tw_1 j = 3h, tw_2 k = 24	10
	Max. features[5]	'auto', 'sqrt'	'auto', 'sqrt'		'auto'
	Min. samples leaf[6]	1, 2, 4	2, 4, 6		5
	Min. samples split[7]	2, 5, 10	2, 3, 4, 5		2
	Number of estimators[8]	100, 200, 300, 400, 500, 600, 700, 800, 900, 1000	100, 500, 800		500
SVM	Kernel[9]	N/A	Linear, rbf, poly, sigmoid	tw_1 j = 24h, tw_2 k = 2	Linear
	C[10]	N/A	0.01, 0.1, 1, 10	tw_1 j = 3h, tw_2 k = 24	0.01
	Degree[11]	N/A	2, 3		2
	Gamma[12]	N/A	'auto', 0.01, 0.1, 1, 10		'auto'

[1] All models used to find the optimum values for each hyperparameter included lying time data, using all sensor data regardless of the time of the day (all day time series).

[2] *k*: number of neighbors.

[3] Bootstrap: method for sampling data points (with or without replacement).

[4] Max. depth: maximum number of levels in each decision tree to control for overfitting.

[5] Max. features: maximum number of features considered for splitting a node.

[6] Min. samples leaf: minimum number of data points allowed in a leaf node.

[7] Min. samples split: minimum number of data points in a node before the node is split.

[8] Number of estimators: number of trees in the forest.

[9] Kernel: type of kernel used to map the data to a different space where a linear hyperplane can be used.

[10] C: cost parameter to control the tradeoff between the misclassifications and width of the margin.

[11] Degree: degree of the polynomial used when kernel = 'poly'.

[12] Gamma: defines how far the influence of a single data point reaches and configures the sensitivity to differences in the data. When gamma is large, the radius of the area of influence only includes the support vector itself, and no amount of regularization with C will be able to prevent overfitting.

Table 3. 2: Descriptive statistics for the five behavioral variables measured with a leg-attached 3-axis accelerometer (TrackaCow, ENGS, Hampshire, UK).

		Time Window[1]				Time of the Day[2]		
	Raw Data	3 h	6 h	12 h	24 h	Milking	Morning	Evening-Night
Lying (minutes per hr.)								
n	14,138	4,374	2,199	1,138	588	2,356	5,312	6,470
mean	21.74	0.33	0.34	0.36	0.46	11.78	21.88	25.25
std	21.06	14.94	10.63	8.39	7.58	15.87	20.81	21.76
min	0	-60	-60	-60	-60	0	0	0
25%	0	-8.67	-5.83	-4.17	-3.05	0	1	2
50%	17	0	-0.17	-0.17	0	3	17	23
75%	39	9.33	6.33	4.56	3.55	19	39	45
max	60	58	56	43	43	60	60	60
Lying Bouts (number per hr.)								
n	14,138	4,374	2,199	1,138	588	2,356	5,312	6,470
mean	0.62	0.00	0.00	0.00	0.00	0.53	0.71	0.57
std	0.85	0.60	0.44	0.33	0.27	0.77	0.94	0.79
min	0	-5	-3.33	-2.67	-2.67	0	0	0
25%	0	-0.33	-0.17	-0.17	-0.13	0	0	0
50%	0	0	0	0	0	0	0	0
75%	1	0.33	0.17	0.17	0.13	1	1	1
max	12	5.33	4.33	2.25	1.5	6	11	12
Steps (number per hr.)								
n	13,631	4,422	2,219	1,142	587	2,372	5,348	5,911
mean	98.80	-1.71	-1.64	-1.77	-1.90	117.64	109.47	81.59
std	72.51	54.52	41.70	34.07	28.53	56.29	81.24	65.88
min	0	-536.50	-291.20	-208.27	-152.33	0	0	0
25%	45	-29	-20.73	-16.95	-13.17	81	49	29
50%	91	-0.67	-0.17	0	-1.30	114	98	73
75%	138	27.67	18.4	14.10	11.61	150	152	118
max	636	448	267.2	202.82	115.70	479	636	574
Intake (min per hr.)								
n	10,874	3,312	1,667	869	452	1,812	4,088	4,974
mean	7.54	0.22	0.22	0.17	0.31	5.30	8.39	7.65
std	12.00	8.69	5.90	4.49	3.46	9.51	12.58	12.23
min	0	-43	-40	-33	-24	0	0	0
25%	0	-4.33	-2.67	-1.83	-1.21	0	0	0
50%	0	0	0.17	0.17	0.35	0	0	0
75%	12	5	3.33	2.42	1.80	7	14	12
max	60	47.33	33.5	25	25	60	60	60

Table 3. 2 (Continued): Descriptive statistics for the five behavioral variables measured with a leg-attached 3-axis accelerometer (TrackaCow, ENGS, Hampshire, UK).

		Time Window[1]				Time of the Day[2]		
	Raw Data	3 h	6 h	12 h	24 h	Milking	Morning	Evening-Night
Intake Visit (number per hr.)								
n	11,330	3,456	1,739	906	471	1,888	4,259	5,183
mean	0.38	0.00	0.00	0.01	0.01	0.39	0.41	0.35
std	0.56	0.37	0.26	0.20	0.17	0.55	0.59	0.54
min	0	-1.67	-1	-0.67	-0.54	0	0	0
25%	0	-0.33	-0.17	-0.08	-0.08	0	0	0
50%	0	0	0	0	0	0	0	0
75%	1	0.33	0.17	0.08	0.08	1	1	1
max	3	1.67	1.17	1	1	3	3	3

[1] Time window: level of hourly sensor data aggregation. Computations were done after removal of seasonality in the raw sensor data by differentiation.

[2] Time of the day: to assess differences based on scheduled farm activities, activities were classified based on farm schedule: milking was from 4:00 to 5:59 h and from 15:00 to 16:59 h; morning was from 6:00 to 14:59 h; evening-night was from 17:00 to 3:59 h of the following day.

Table 3. 3: Results from models performance (%) where random forest (RF) classifier was used on sensor data registered by a leg-attached 3-axis accelerometer (TrackaCow, ENGS, Hampshire, UK) from all day were aggregated using a 6 hour time window. Different cut-off values were chosen using classification probabilities ranked from high to low.

Behavior	Time Lag *k*	Sample Size		20% cut-off						30% cut-off					
		Metritis	Non-metritis	Se	Sp	PPV	NPV	Ac.	F_1 score	Se	Sp	PPV	NPV	Ac.	F_1 score
Lying	7	43	163	93.02	99.39	97.56	98.18	98.06	95.24	100	88.96	70.49	100	91.26	82.69
	8	43	162	93.02	99.38	97.56	98.17	98.05	95.24	100	88.89	70.49	100	91.22	82.69
	9	42	161	90.48	98.76	95	97.55	97.04	92.68	100	88.82	70	100	91.13	82.35
	10	41	156	92.68	99.36	97.44	98.1	97.97	95	100	88.46	69.49	100	90.86	82
	11	41	156	92.68	99.36	97.44	98.1	97.97	95	100	88.46	69.49	100	90.86	82
	12	40	155	92.5	98.71	94.87	98.08	97.44	93.67	100	88.39	68.97	100	90.77	81.64
Step	7	45	167	91.11	99.4	97.62	97.65	97.64	94.25	100	89.22	71.43	100	91.51	83.33
	8	45	166	93.33	100	100	98.22	98.58	96.55	100	89.16	71.43	100	91.47	83.33
	9	44	165	90.91	99.39	97.56	97.62	97.61	94.12	100	89.09	70.97	100	91.39	83.02
	10	44	160	88.64	99.38	97.5	96.95	97.06	92.86	100	89.38	72.13	100	91.67	83.81
	11	44	160	88.64	99.38	97.5	96.95	97.06	92.86	100	89.38	72.13	100	91.67	83.81
	12	43	159	88.37	98.74	95	96.91	96.53	91.57	100	89.31	71.67	100	91.58	83.50

Table 3. 4: Results from models' performance (%) were random forest (RF) classifier was used on sensor data registered by a leg-attached 3-axis accelerometer (TrackaCow, ENGS, Hampshire, UK) from all day aggregated using time windows of 3 hours. Different cut-off values were chosen using classification probabilities ranked from high to low. Only rows where a change in either sensitivity (Se) or positive predictive value (PPV) at the 30% cut-off are shown.

Behavior	Time Lag *k*	Sample Size		20% cut-off						30% cut-off					
		Metritis	Non-metritis	Se	Sp	PPV	NPV	Ac	F_1 score	Se	Sp	PPV	NPV	Ac	F_1 score
Lying	6	44	168	93.18	99.4	97.62	98.24	98.11	95.35	100	88.69	69.84	100	91.04	82.24
	8	44	168	95.45	100	100	98.82	99.06	97.67	100	88.69	69.84	100	91.04	82.24
	10	44	164	93.18	100	100	98.2	98.56	96.47	100	89.02	70.97	100	91.35	83.02
	11	43	159	93.02	100	100	98.15	98.51	96.38	100	89.31	71.67	100	91.58	83.5
	16	42	157	92.86	100	100	98.13	98.49	96.3	100	89.17	71.19	100	91.46	83.17
	23	40	150	95	100	100	98.68	98.95	97.44	100	88.67	70.18	100	91.05	82.48
	24	40	149	92.5	100	100	98.03	98.41	96.1	100	89.26	71.43	100	91.53	83.33
Step	6	45	171	88.89	98.25	93.02	97.11	96.3	90.91	100	88.89	70.31	100	81.02	68.71
	7	45	171	91.11	98.83	95.35	97.69	97.22	93.18	100	88.89	70.31	100	81.02	68.71
	10	45	167	91.11	99.4	97.62	97.65	97.64	94.25	100	89.22	71.43	100	81.6	69.77
	11	45	163	91.11	100	100	97.6	98.08	95.35	100	89.57	72.58	100	81.73	70.32
	12	45	163	88.89	99.39	97.56	97.01	97.12	93.02	100	89.57	72.58	100	81.73	70.32
	14	45	163	91.11	100	100	97.6	98.08	95.35	100	89.57	72.58	100	81.73	70.32
	16	44	161	93.18	100	100	98.17	98.54	96.47	100	89.44	72.13	100	81.46	69.84
	18	44	159	88.64	99.37	97.5	96.93	97.04	92.86	100	89.94	73.33	100	81.77	70.4
	19	44	155	88.64	100	100	96.88	97.49	93.98	100	90.32	74.58	100	82.41	71.55
	22	44	155	86.36	99.35	97.44	96.25	96.48	91.57	100	90.32	74.58	100	82.41	71.55
	23	43	154	88.37	99.35	97.44	96.84	96.95	92.68	100	89.61	72.88	100	82.23	71.08
	24	43	153	90.7	100	100	97.45	97.96	95.12	100	90.2	74.14	100	82.14	71.08

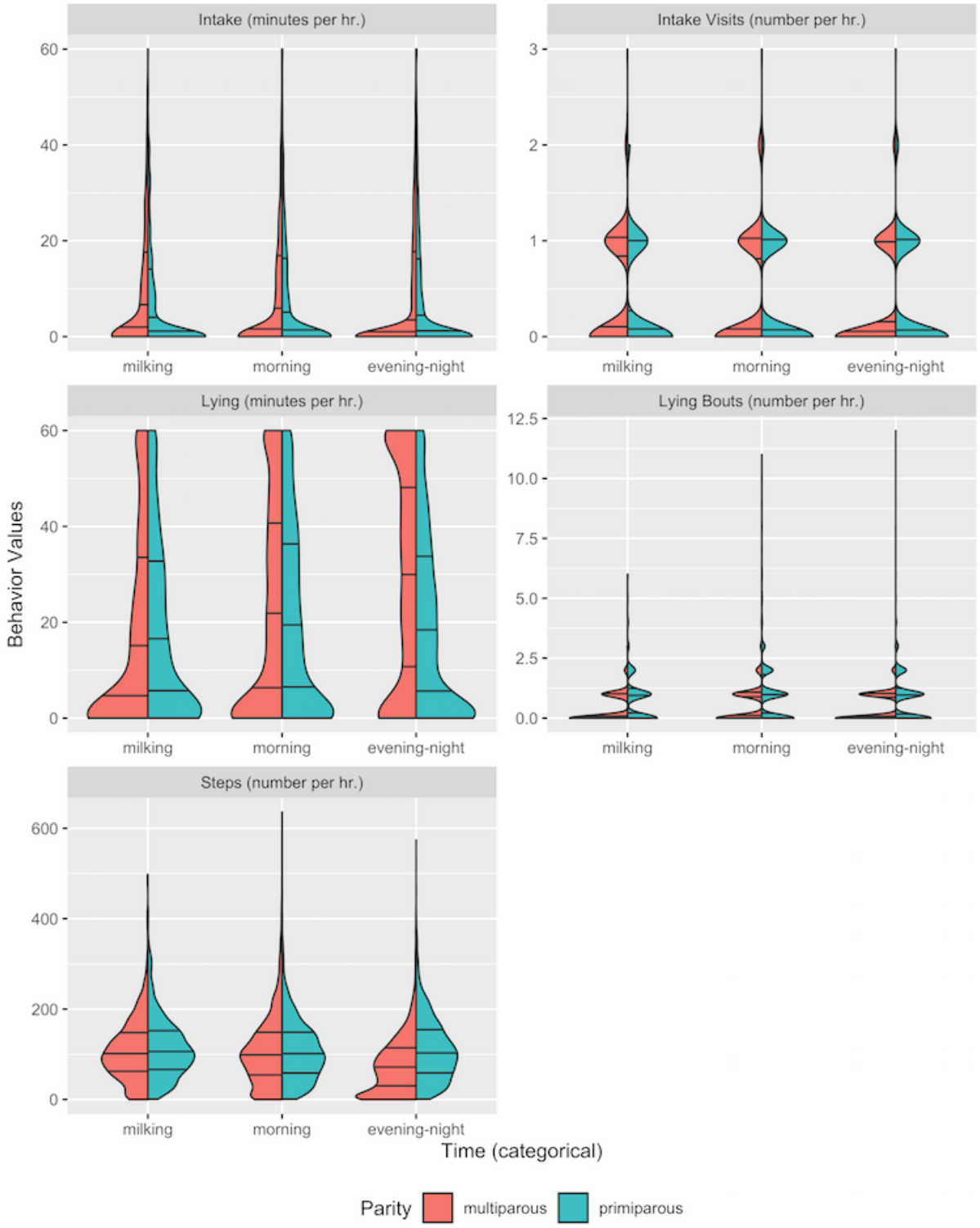

Figure 3. 1: Distribution and density of raw sensor data stratified by parity and time of the day for the five behaviors registered by a leg-attached 3-axis accelerometer (TrackaCow, ENGS, Hampshire, UK). Horizontal lines indicate mean and standard deviation. Milking is frm 4:00 to 5:59 h and from 15:00 to 16:59 h; morning is from 6:00 to 14:59 h; evening-night is from 17:00 to 3:59 h in the following day.

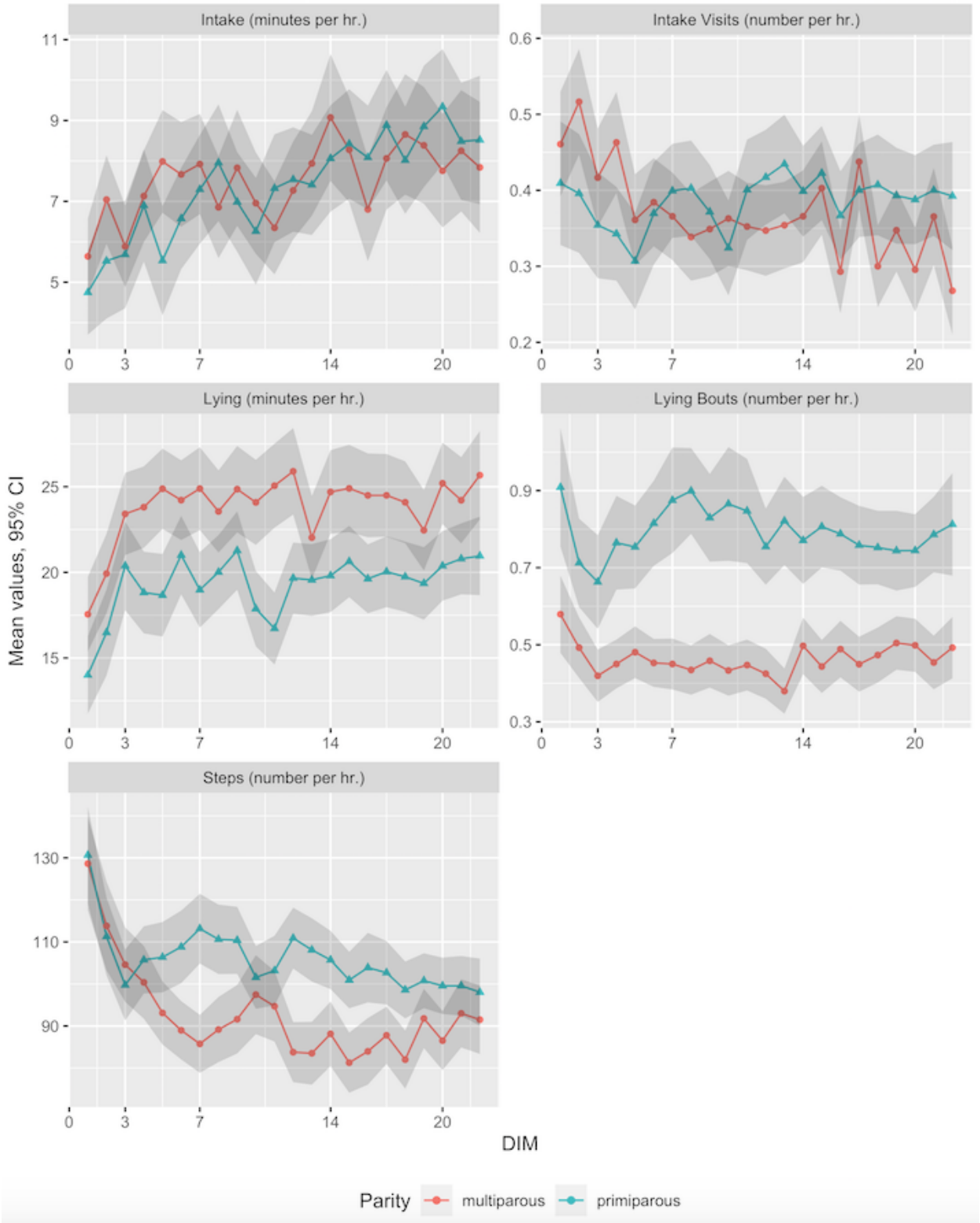

Figure 3. 2: Mean raw sensor data and 95% C.I. for the mean by days in milk (DIM) stratified by parity for the five behavioral variables registered by a leg-attached 3-axis accelerometer (TrackaCow, ENGS, Hampshire, UK) for the whole study period.

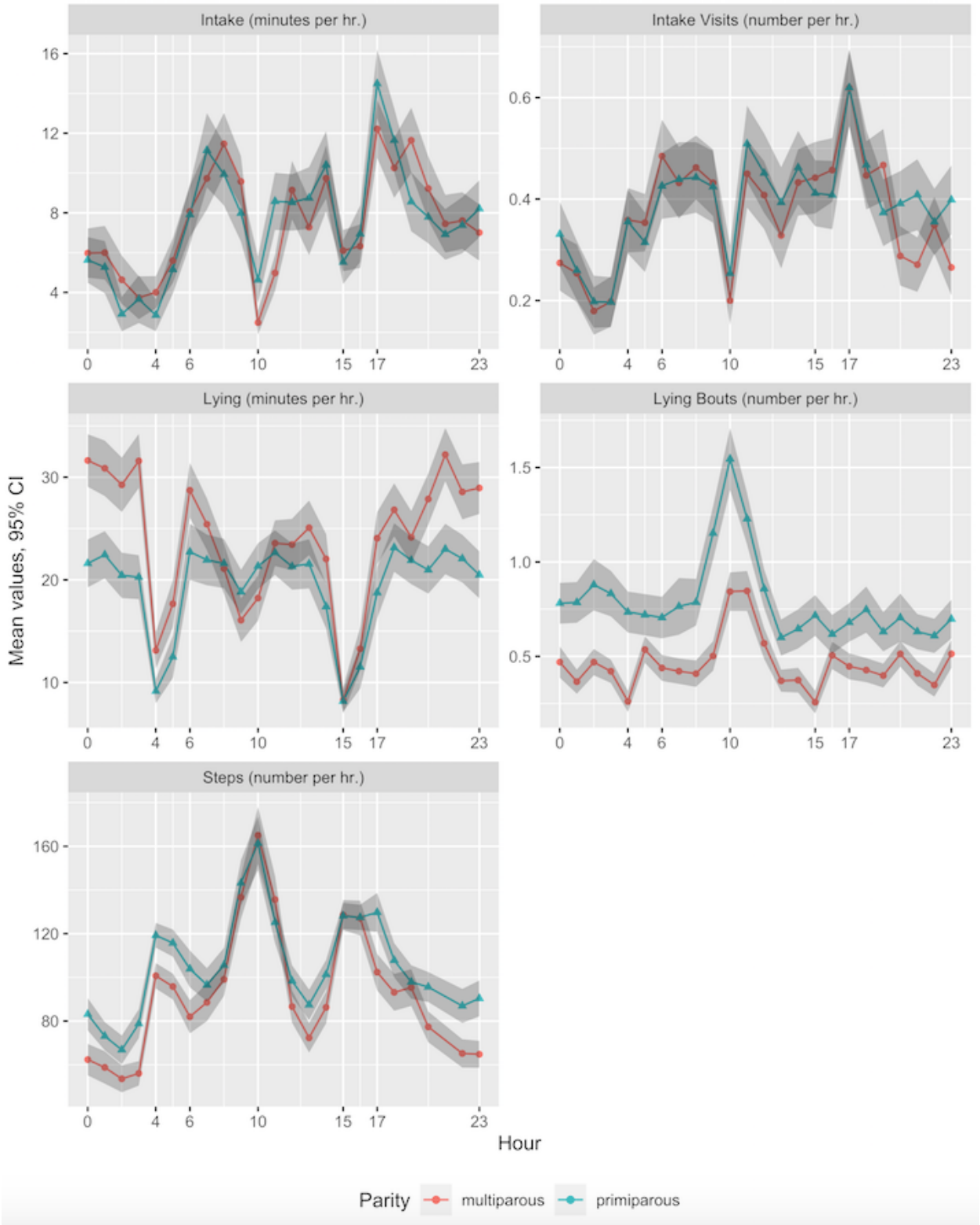

Figure 3. 3: Mean raw sensor data and 95% C.I. for the mean by the time of the day (Hour) stratified by parity for the 5 behavioral variables measured with a leg-attached 3-axis accelerometer (TrackaCow, ENGS, Hampshire, UK).

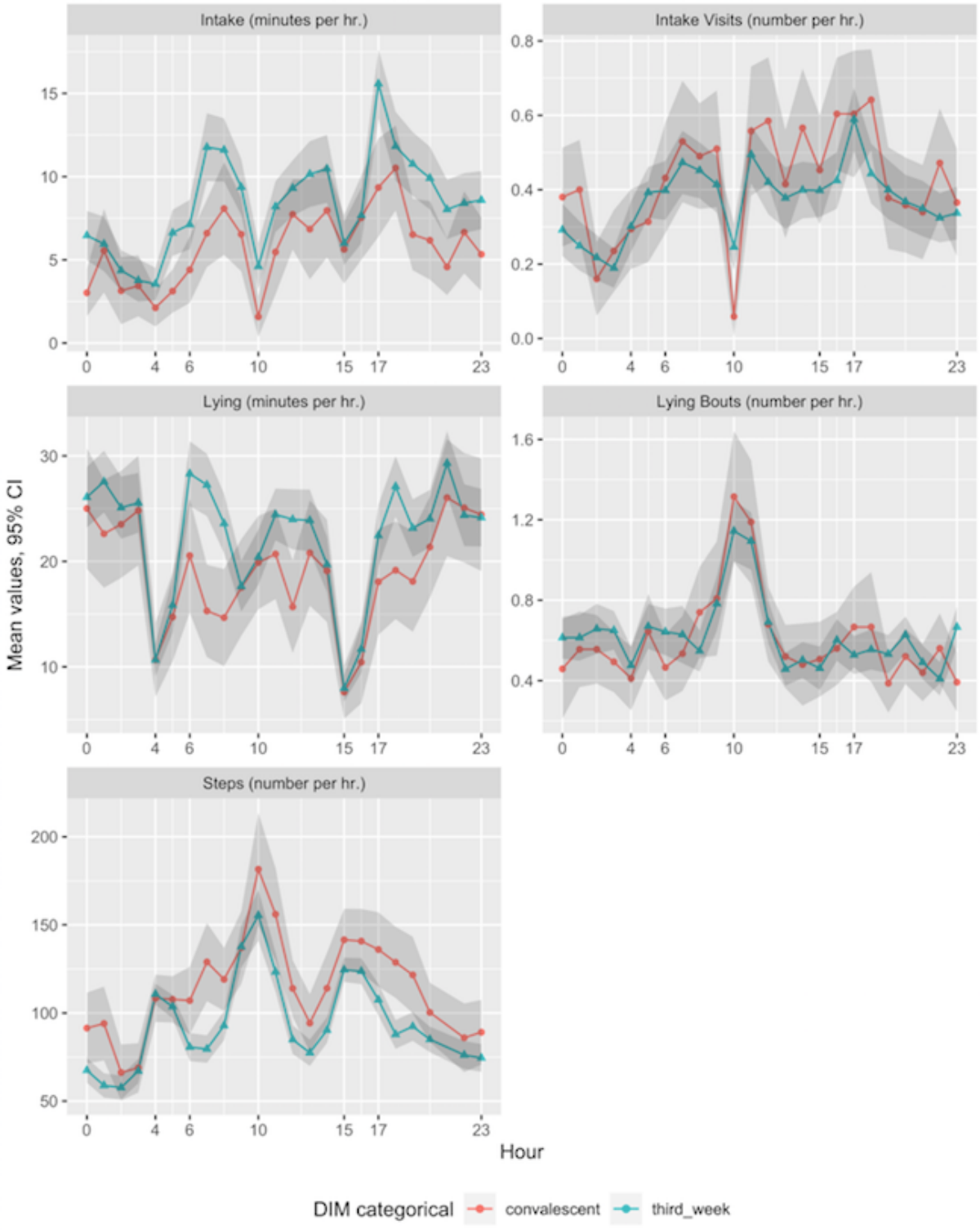

Figure 3. 4: Mean raw sensor data and 95% C.I. for the mean for each behavioral variable measured with a leg-attached 3-axis accelerometer (TrackaCow, ENGS, Hampshire, UK) in a 24 hour period stratified by parity and days in milk (DIM) categorized as convalescent (parturition to 3 DIM), first week (4 – 7 DIM), second week (8 – 14 DIM), and third week (15 – 21 DIM).

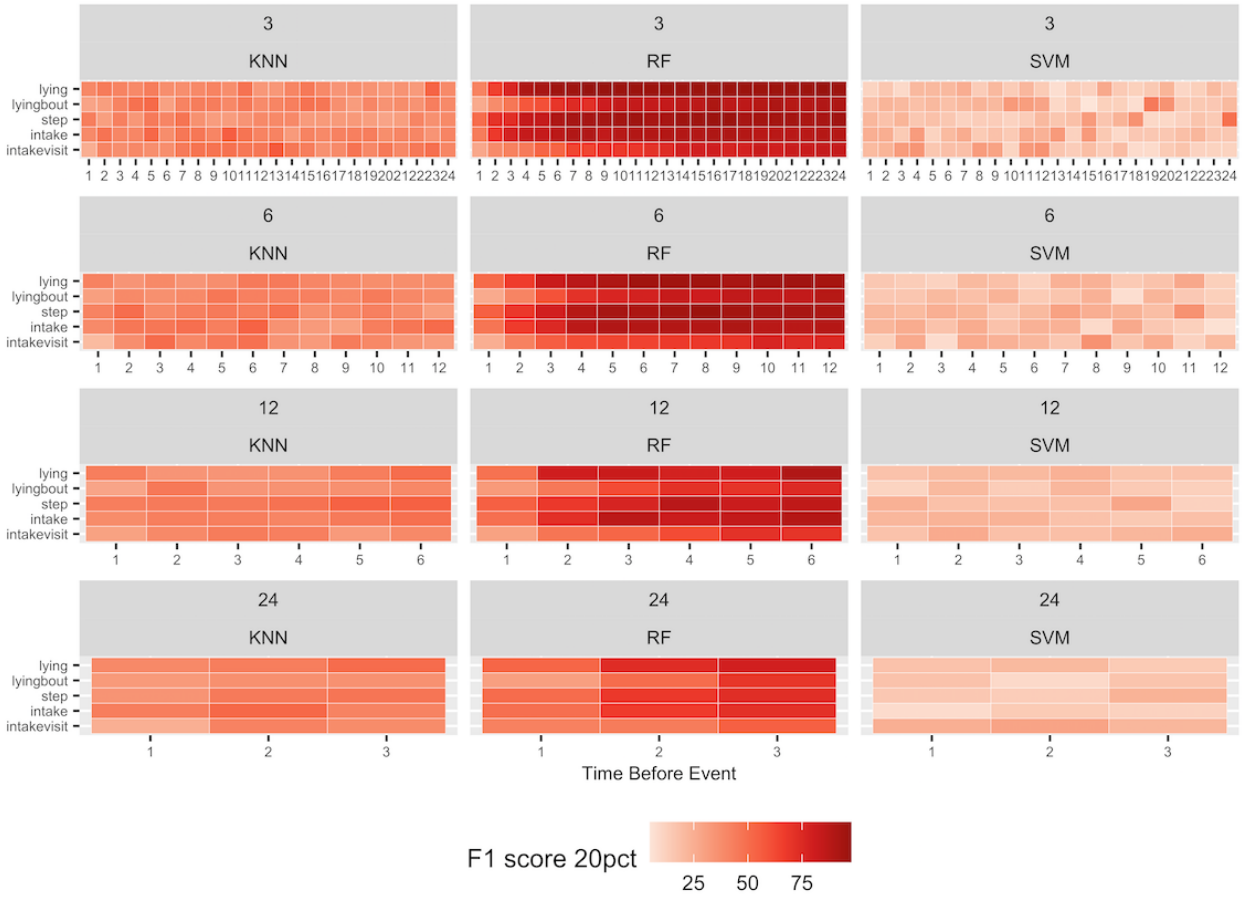

Figure 3. 5: F_1 scores (%) using the 20% highest class probabilities as cut-off when sensor data registered by a leg-attached 3-axis accelerometer (TrackaCow, ENGS, Hampshire, UK) were aggregated using time windows of 24, 12, 6, and 3 hours. F_1 scores are shown for those models where all sensor data were used to fit the models and parity was not taken into account. F_1 scores are shown for different time lags and for each one of the classifiers (*k*-nearest neighbors, random forest, and support vector machines).

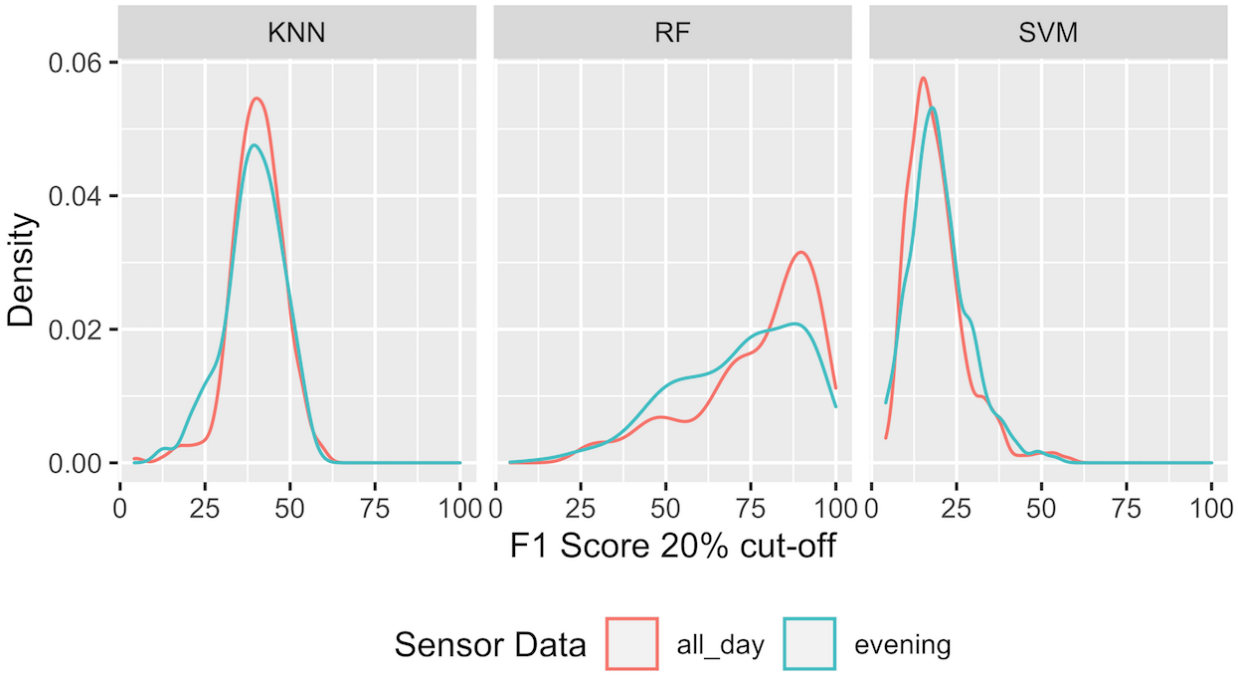

Figure 3. 6: Distribution of F_1 scores (%) using the 20% highest class probabilities as threshold when sensor data registered by a leg-attached 3-axis accelerometer (TrackaCow, ENGS, Hampshire, UK) were aggregated using time windows of 24, 12, 6, and 3 hours.

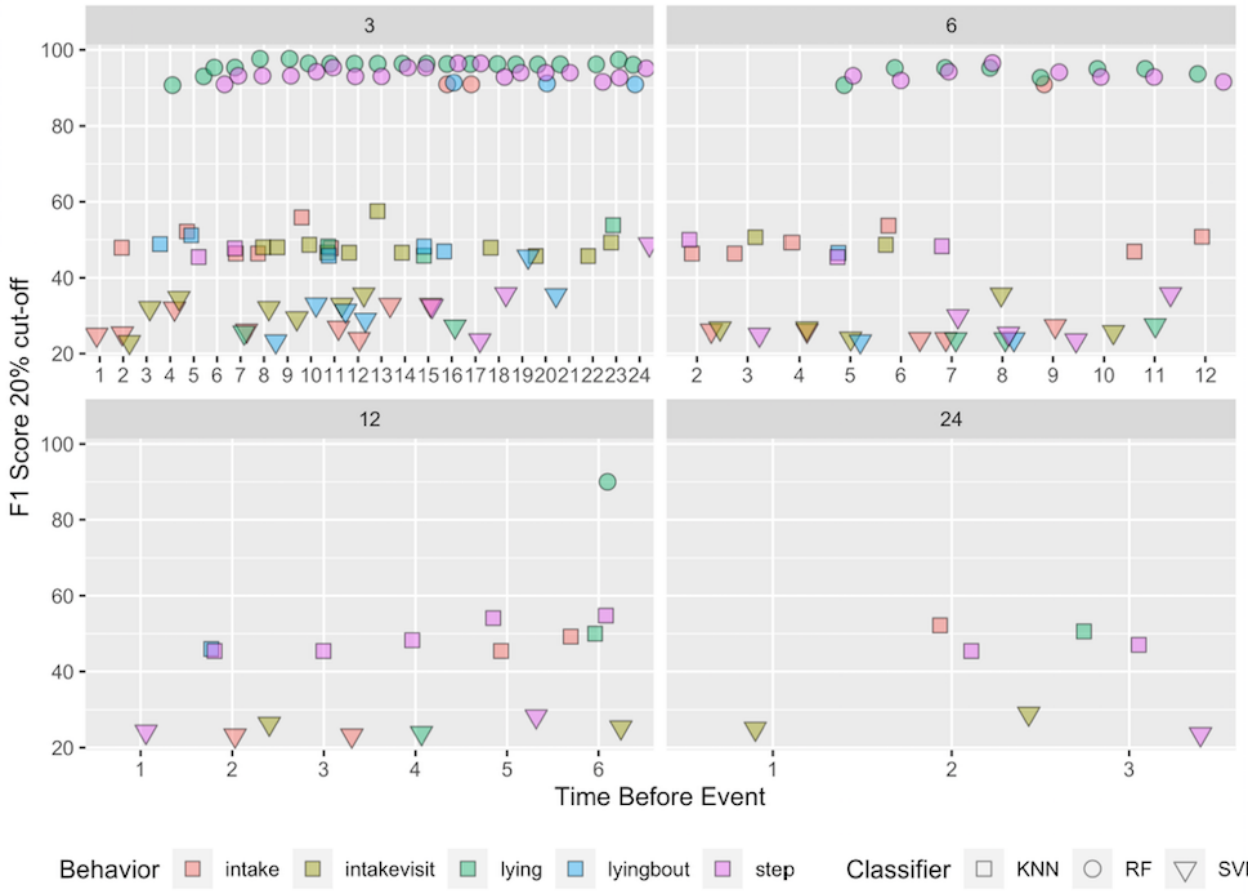

Figure 3. 7: Distribution of F_1 scores (%) at the 20% cut-off from the upper quartile by behavior and classifier when sensor data registered by a leg-attached 3-axis accelerometer (TrackaCow, ENGS, Hampshire, UK) were aggregated using 24, 12, 6, and 3 time windows, and sensor data from all day were used.

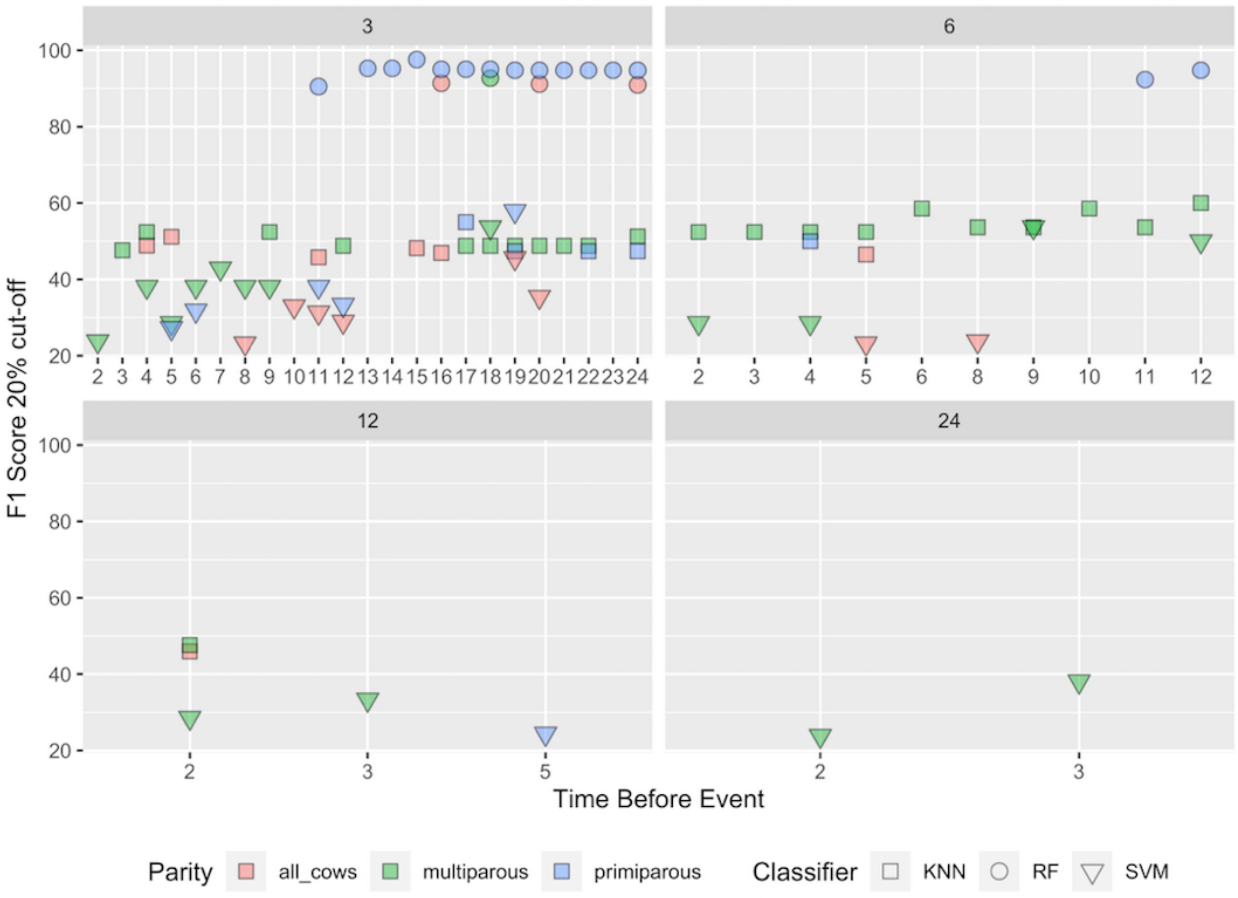

Figure 3. 8: Distribution of lying bouts F_1 scores (%) at the 20% cut-off by classification algorithm and parity stratified by different time windows (3 and 6 hours).

4 Systematic approach to performance analysis of a machine learning classifier to predict metritis events in dairy cattle using multiple data streams.

G. Vidal,[1] J. Sharpnack,[2] P. Pinedo,[3] I. C. Tsai,[4] A. R. Lee,[4] and B. Martínez-López[1]

[1] Center of Animal Disease Modeling and Surveillance (CADMS). Department of Medicine and Epidemiology. School of Veterinary Medicine. University of California Davis, Davis, CA, 95616.

[2] Department of Statistics. University of California Davis, Davis, CA 95616.

[3] Department of Animal Sciences. Colorado State University, Fort Collins, CO, 80523.

[4] Department of Animal Sciences. College of Agriculture, Food, and Environment. University of Kentucky, Lexington, KY, 40546.

4.1 Abstract

Due to the large volume of data available, there is the need to study machine learning algorithms performance at different time windows and time lags when inputs from multiple behaviors are included simultaneously in the model, and to combine those with other data inputs such as milk yield-related variables. Our objectives were 1) compare the classifier performance when, for a given device, multiple behavioral patterns are combined using a model selection framework; 2) study whether classifier performance improves when multiple behavioral patterns measured with two different devices are combined; 3) study whether classifier performance improves when milk-related variables are added to the best selected models; and 4) compare model performance under nowcasting and forecasting frameworks. Data from 35 dairy cows that either did not experience any disease postpartum or were only affected by metritis were retrospectively selected from a dataset containing sensor data and clinical data from 138 lactating cows during the first 21 days postpartum at University of Kentucky Coldstream Dairy from June 2014 to May 2017. Metritis events were created based on changes in metritis scores recorded during clinical examination. Random forest was used on sensor data from ten different cow behaviors that were aggregated using 12, 6, and 3 hour time windows. Better results were obtained when a 20% threshold was used to classify the observations into cases and no-cases. At shorter time lags, performance decreased across all time windows, with model complexity increasing in order to maintain performance levels. The addition of milk-yield related variables did not always improved performance, and in those cases where it did, it was driven by an increase in sensitivity. Furthermore, the combination of devices did not improve classification performance. Lastly, forecasting models had a level of

performance comparable to those resulted from nowcasting models, with the advantage that earlier interventions could be implemented.

Keywords: predictive modeling; random forest; sensor fusion; forecasting; nowcasting

4.2 Introduction

Different health states have an impact on animal welfare and economic efficiency of dairy farms. During the post-partum period, their impact is associated with decreased milk production (Fourichon et al., 1999; Edwards and Tozer, 2004; Huzzey et al., 2007b), poor reproductive performance (Opsomer et al., 2000; Walsh et al., 2007), and increased culling rate (Dubuc et al., 2011). Therefore, early detection of sick animals can have an impact on productivity and welfare if medical interventions are implemented in a timely manner. The use of existing commercial sensors combined with data-driven modeling approaches can aid in the detection of sick animals in real-time based on changes in high frequency and low frequency data such as behavioral patterns, daily milk yield, or parity (Steensels et al., 2016).

There is a growing body of literature dedicated to the validation of changes in behaviors measured by precision dairy farming technologies (PDFT) for heat and disease detection such as ketosis, metritis, mastitis, or lameness under different management systems such as freestall (Dolecheck et al., 2015) or pasture (Kamphuis et al., 2012; Sepúlveda-Varas et al., 2014). Different animal behaviors such as rumination, eating, lying, lying bouts, steps, and activity are usually recorded with PDFTs, and data-driven modeling approaches such as machine learning classifiers for binary outcomes (healthy - sick) are commonly used for predictive modeling. For disease prediction, common classifiers are decision trees (DT) (Kamphuis et al., 2010a; Steensels et al., 2016; Tamura et al., 2019), random forest (RF) (Vidal et al., Chapter 1 and 2), support vector machines (SVM) (Vanrell et al., 2014), and logistic regression models (LR) with or without random effects (Urton et al., 2005). Overall, decision tree-based methods such as random forest,

have yielded better performance when compared to other methods (Vidal et al., Chapter 1 and 2).

Monitoring a wider set of behaviors has been hypothesized to be of greater predictive value for detecting sick animals compared with more restricted set of behaviors (Matthews et al., 2016, 2017). However, despite the high number of behaviors being recorded by PDFTs, these are rarely combined and only a limited number of behavioral variables are usually included in prediction models (Saint-Dizier and Chastant-Maillard, 2018). The most common classifier performance metrics used in PDFT literature are sensitivity (Se), specificity (Sp), positive and negative predictive values (PPV, NPV), accuracy (Ac), and F_1 score. However, results across studies differ due to differences in methodologies, time windows used to aggregate sensor data, time lags, and metrics chosen, making the comparison across studies difficult. Despite the fact that classification performance is affected by the time lags chosen (Saint-Dizier and Chastant-Maillard, 2018), changes in classification performance due to low prevalence and sensor sampling strategies to improve performance have been ignored in PDFT literature (Carslake et al., 2021). Similarly, the study of the impact of different time windows has been poorly studied in the PDFT literature when trying to predict animal health with animal behavior, as many studies fail to establish which signal features and sampling rates are most appropriate for each behavior (Carslake et al., 2021). Therefore, there is the need to study algorithms performance at different time windows and time lags when inputs from multiple behaviors are included simultaneously in the model, and to combine those with other data inputs such as milk yield-related variables.

The objectives of the present study were: 1) to compare classifier performance when, for a given device, multiple behavioral patterns are combined using a model selection framework; 2)

study whether classifier performance improves when multiple behavioral patterns measured with two different devices are combined; 3) study whether classifier performance improves when milk yield-related variables are added to the best selected models; and 4) compare model performance under nowcasting and forecasting working frameworks. Our findings will be needed for the development of cost analysis models where the cost of the investment of different devices, data storage and data analysis, and cost of medical intervention in terms of labor, duration of medical interventions, and milk revenues could be evaluated.

4.3 Material and Methods

The data used in this study was part of a large study designed to quantify physiological and behavioral changes associated with mastitis, lameness, estrus, and postpartum diseases, using multiple PDF technologies (Tsai, 2017; Lee, 2018). The larger study included data from 138 lactating cows at the University of Kentucky Coldstream Dairy (Lexington, KY, USA) that were enrolled in the study during two study periods (June 2014 to October 2015 and July 2016 to May 2017). All studies were performed with the approval of the University of Kentucky Institutional Animal Care and Use Committee (IACUC protocol number 2013-1199 and 2016-2368).

4.3.1 Population Data

From the original dataset, a total of 35 dairy cows (Jersey = 20; Holstein = 15) that either did not experience any disease postpartum or were only affected by metritis were retrospectively selected. Cows were enrolled in the study after parturition and were followed

during 21 days postpartum, and were removed from the study if they died or were culled from the herd before 21 days in milk (DIM).

Details about farm management are described elsewhere (Tsai, 2017; Lee, 2018). Briefly, about one month before expected calving date, cows were moved from a far-off dry pen and pasture to a close-up dry pen. Cows were maintained in a fresh cow pen from parturition to 70 DIM. Subsequently, lactating cows were housed in two freestall barns. During the first study period, one barn had 54 dual chamber waterbeds (Advanced Comfort technology, Inc., Reedsburg, WI) and the other was equipped with 54 rubber-filled mattresses, both surfaces covered with sawdust. During the second study period, both barns had compost bedded pack tilled twice daily, and bedded with sawdust as needed. Cows were provided *ad libitum* access to fresh water in each barn and lactating cows were fed the same TMR between 6:00 to 9:30 h and 12:30 to 15:00 h. The lactating diet consisted of forage, alfalfa hay, mineral and vitamin supplement, concentrate mix, whole cottonseed, and alfalfa haylage. During the second study period, feed was pushed up 22 times per day by an automated feed pusher (Lely Juno, Ley Robots, Masslius, the Netherlands). Cows were milked two times per day from 4:30 to 5:30 h and from 15:30 to 16:30 h in a double 2 X 2 tandem-milking parlor.

4.3.2 Clinical Data

Fresh cows were monitored daily after morning milking from 7:00 to 10:00 h for the first 21 days of lactation. A MetriCheck (Simero Tech Ltd, Hamilton, New Zealand) device was used to obtain a uterine discharge sample and scored on 3, 5, 7, 9, 11, 17, 19, and 21 DIM. Depending on the study period, different number of uterine discharge samples were taken between 11 and 17

DIM: during the first study period, one sample was taken on 14 DIM, while during the second study period, samples were taken on 13 and 15 DIM. Each uterine discharge was evaluated on a 1 to 3 scale modified from Sheldon et al. (2006). Briefly, score 1: thick, viscous discharge, clear, opaque or red to brown in color, no odor or milk; score 2: white or yellow pus, moderate to thick discharge, milk odor; score 3: pink, red, dark red, or black watery discharge, detectable offensive odor, possibly intolerable. Cows with score $\geq$ 2 were classified as metritis cases (Tsai, 2017; Lee, 2018). As part of the study, cows were monitored for hyperketonemia, hypocalcemia, mastitis, lameness, and retained placenta as described by Tsai (2017) and Lee (2018). Briefly, blood was collected by caudal venipuncture on 3, 7, 14, and 21 DIM for calcium level from blood serum and non-esterified fatty acid determination (NEFA), while beta-hydroxybutyrate (BHBA) concentration was measured with two cowside monitors. Precision Xtra (Abbott Laboratories, Chicago, IL, USA) was used on days 3, 7, 14, and 21 post-partum during the first study period, while BHBCheck (PortaCheck Inc., Moorestown NJ, USA) was used on days 1, 2, 3, 4, 5, 6, 7, 10, 14, and 21 post-partum during the second study period. Hypocalcemia was defined as a serum Ca level <8.6 mg/dL (Chapinal et al., 2011) and hyperketonemia was diagnosed when BHBA $\geq$ 1.2 mmol/L (Geishauser et al., 1998; McArt et al., 2012; Kaufman et al., 2016). Cows were diagnosed with clinical mastitis using the following criteria: watery, thickened, and discolored milk; milk containing blood, pus, flakes, or clots; edema, erythema; or pain in the associated quarter (Royster and Wagner, 2015) between 1 and 21 DIM by trained milkers. Furthermore, quarter milk samples were collected for somatic cell count (SCC) on days 4 +/- 2 DIM and 9 +/- 2 DIM. Cows with SCC $\geq$ 200,000 cells/mL in one or more quarters were considered positive for subclinical mastitis. Finally, locomotion scores were recorded on days 1, 7, 14, and 21

postpartum on a 1 to 3 scale (Schlageter-Tello et al., 2014). Retained placenta was recorded if fetal membranes were retained for > 24 hours (Sheldon et al., 2006).

For any given cow and day, a metritis event was assigned when a cow was getting or being with metritis, this is, the metritis score increased, changed from 3 to 2, or when the score remained 2 or 3, between two consecutive uterine discharge evaluations. Similarly, for any given cow and any given day, a non-metritis event was assigned when a cow was recovering from metritis or being healthy, this is, when the metritis score decreased to 1, or when the score remained as 1, between two consecutive uterine discharge evaluations. In order to keep the time relationship between sensor measurements and clinical data, diagnosis of metritis was assigned to happen at 6:00 h on each one of the days when uterine discharge was evaluated.

4.3.3 Sensor Data and Data Pre-processing

Each cow was equipped with different PDM technologies before being enrolled to allow for an adjustment period of at least two weeks. For this retrospective study, information per cow included ten different behaviors measured from parturition to 21 days postpartum with a 3-axis accelerometer attached to the ear (CowManager, Agis Autimatisering, Harmelen, Netherlands), and a 3-axis accelerometer attached to the leg (TrackaCow, ENGS; Hampshire, UK). The ear tag device records the number of minutes per hour for behaviors classified as rumination, eating, not active (this could happen either while standing or lying), active, or high activity. CowManager has been previously validated by Bikker et al. (2014) and Borchers et al. (2016). The leg-attached accelerometer records hourly data for lying time (minutes per hour), lying bouts (number per hour), steps (number per hour), intake (minutes per hour), and intake visit (number of visits to

the feedbunk per hour). TrackaCow has been previously validated by Borchers (2015), Borchers et al. (2016), and Chapinal et al., (2007).

Time series sensor data consisted on the hourly measurements for each behavior i corresponding to the 3 days prior to each metritis event, assigning the time of diagnosis t at 6:00 h on each one of the days when uterine discharge was evaluated. Therefore, the 6:00 h time was used as offset for later transformations of the time series sensor data.

The first time series data transformation was to remove seasonality by differencing the time series. In order to do that, we subtracted for each cow, behavior i, and hour within a 3 day period before a given metritis event, the measurement registered by the sensor in the previous 24 h from each hourly sensor measurement. The time series data for each metritis event at time t was defined by:

$$\left(x_{i,t-1}, x_{i,t-2}, \ldots, x_{i,t-n}\right)$$

where:

x_i was the differenced hourly sensor measurement for behavior i and time t, being $i \in \{rumination, eating, not\ active, active, high\ activity, lying, lying\ bouts, steps, intake, intake\ visit\}$

n was the time step within a 3 day (or 72 hours) period.

Next, we transformed the time series sensor data by aggregating the differenced hourly measurements using the mean of the time window tw_1. In order to assess classifier performance at different levels of sensor data aggregation, we used different widths for the time window tw_1 to compute the mean. Based on our previous findings, classifier performance changes depending on the device and width of the time window tw_1. Therefore, the time series generated by CowManager device was aggregated by computing the mean using widths of 6 and 12 hour for

the time window tw_1 (Vidal et al., Chapter 1). In contrast, the time series generated by TrackaCow device was aggregated by computing the mean using widths of 3 and 6 hour for the time window tw_1 (Vidal et al., Chapter 2). As result, the new time series data for each metritis event at time t was defined by:

$$\left(\bar{x}_{ij,t-1}, \bar{x}_{ij,t-2}, \ldots, \bar{x}_{ij,t-m}\right)$$

where:

$\bar{x}_{ij}$ was the mean sensor value for behavior i and time window tw_1 of width j,

being $i \in \{rumination, eating, not\ active, active, high\ activity, lying, lying\ bouts, steps, intake, intake\ visit\}$

and $j \in \{3\ h, 6\ h, 12\ h\}$, depending on the device.

m was the time step within a 3 day period. The number of time steps that could be included within a 3 day period was a function of the width j of the time window tw_1.

4.3.4 Model Building

Model building was done under two different frameworks: nowcasting and forecasting, and the number of time steps included as features changed depending on the framework used.

- <u>Nowcasting Framework:</u>

  Under a nowcasting framework, the selection of model features was conducted by using a time window tw_2 of width k. Furthermore, in order to assess classifier performance at different widths, we used multiple values for k within a 3 day period before each metritis event. Therefore, the model inputs for the classifier were:

$$\left(\bar{x}_{ij,t-1}, \bar{x}_{ij,t-2}, \ldots, \bar{x}_{ij,t-k}\right)$$

where the width $k = 1, 2, \dots, l$, and l number of time steps included as features within a 3 day period before a given metritis event. The number of features in our models ranged from 1, when sensor data were aggregated with a tw_1 width j of 24 hours and tw_2 width k of 1, to 24 features when sensor data were aggregated using a tw_1 width j of 3 hours and tw_2 width k of 24, corresponding to 72 hours prior to the event.

- Forecasting Framework:

  Under a forecasting framework, the selection of model features was conducted by using a time window tw_2 of width k and taking into account the number of time steps ahead q for our predictions corresponding to 2 or 3 days ahead for a given event. Therefore, the model inputs for the classifier were:

$$(\bar{x}_{ij,t-q}, \bar{x}_{ij,t-q-1}, \bar{x}_{ij,t-q-2}, \dots, \bar{x}_{ij,t-k})$$

  where the width $k = q + 1, q + 2, \dots, l$, being the number of time steps skipped q a function of the width j of time window tw_1 and number of days ahead, and l the number of time steps included as features within a 24 hour period. Therefore, for width $j = 3\ h$, then $q = 8$ or $q = 16$ for the 2 or 3 day forward, respectively. Similarly, for width $j = 6\ h$, then $q = 4$ or $q = 8$ for the 2 or 3 day forward, respectively. Lastly, for width $j = 12\ h$,then $q = 2$ or $q = 4$ for the 2 or 3 day forward, respectively.

Next, a series of modeling steps were used under nowcasting and forecasting frameworks in order to build the models, starting with those that included only features from one behavior at a time, following by those that included features from multiple behaviors, and finishing by adding milk yield-related variables as features to see whether classification performance could improve. Under each framework, the following modeling steps were implemented:

- **Step 1.** One model was fitted for each combination of behavior i and time window tw_1 of width j, and time window tw_2 of width k within a 3 day period. Resulting models were ranked from greatest to smallest F_1 score;
- **Step 2.** The two models with greatest F_1 score from modeling step 1 were selected, and a model that combined each one of their features was fitted. If the F_1 score of the resulting model increased, the next best model from step 1 was selected and its features added into the model. If the F_1 score of the resulting model did not increase, the second best model from step 1 was removed and the features from the next best model from step 1 were added into the model. Best model (greatest F_1 score, Se, and PPV) was selected from step 1 and 2. Among competing models (equal Se, PPV, and F_1 score), models with fewer number of features were preferred.
- **Step 3.** Next, milk yield-related features were added to the best model from previous steps. The milk-related features used were: daily milk yield (m; kg) from each one of the 3 days before an event as model inputs $(m_{t-1}, m_{t-2}, m_{t-3})$, the mean milk yield of the last 3 days before event as model input $\left(\sum_{p=1}^{3} m_{t-p}/3\right)$, the variance of the milk yield of the last 3 days before the event as model input $\left(\sum_{p=1}^{3} m_{t-p}^2 - \bar{m}^2\right)$, the slope of the milk yield of the last 3 days before the event as model input $(m_{max} - m_{min}/2)$, and the combination of daily milk yield (kg) and milk yield variance as model inputs $\left(m_{t-1}, m_{t-2}, m_{t-3}, \sum_{p=1}^{3} m_{t-p}^2 - \bar{m}^2\right)$. Under forecasting framework only the milk yield (kg) corresponding to day -2 or day -3 before event was used for the 2 and 3 days periods, respectively.

4.3.5 Model Performance

Based on our findings from previous studies, a RF classifier was implemented to fit the models (Vidal et al., Chapter 1 and 2). Due to limitations in the amount of data available, we used fivefold cross-validation (5-FCV) to set aside a validation set and use it to assess the performance of the prediction model. Randomized search (RS) was used as strategy to optimize the classifier. Optimal parameters that were found to allow for best mean cross-validation accuracy were used to train the final model at each modeling step. After optimization, we used a rank-based method to classify the events, where the prediction class probability for each health event of being classified as metritis was ranked from highest to lowest. To estimate the performance of each model, highest 20, 30, and 40% class probabilities were used as different thresholds. For each threshold, classification performance was evaluated using estimates of sensitivity (Se or recall), specificity (Sp), positive predictive value (PPV or precision), negative predictive value (NPV), accuracy (Ac), F_1 score, the area under the curve (AUC) for the receiver operating characteristic (ROC) curve, and Precision Recall (PR)-curves. Sensitivity was estimated as the ratio of correctly predicted positive observations to all observations in the actual class (metritis event). Specificity was estimated as the ratio of correctly predicted negative observations to all observations in the actual class (non-metritis event). Positive predictive value was the ratio of correctly predicted positive observations to all predicted positive observations. Similarly, NPV was the ratio of correctly predicted negative observations to all predicted negative observations. Accuracy was the ratio of correct predictions to all number of observations (Hogeveen et al., 2010). F_1 score is the weighted average of PPV and Se and, therefore, it is preferred over Ac in situations where the dataset is unbalanced (Davis and Goadrich, 2006; Saito and Rehmsmeier, 2015). F_1 score was

computed as $(1 + b^2)*(PPV* Se)/((b^2 * PPV) + Se)$, where b = 1 (Saito and Rehmsmeier, 2015). At each modeling step, best performance was defined as the greatest possible values for F_1 score, Se, and PPV, in that particular order.

Classifier implementations were taken from the open-source Python library scikit-learn (Pedregosa et al., 2011). The feature extraction and the optimization of the classifier parameters were implemented using Python programming language, version 2.7 (Python Software Foundation, http://www.python.org). Plots were done using ggplot2 library (Wickham, 2009), using R open-source statistical software (R Core Team, 2017).

4.4 Results

A total of 35 dairy cows (primiparous = 17; multiparous = 18) were retrospectively selected from the original dataset (n = 138) containing clinical and sensor data from parturition to 21 DIM. The average (± SD) milk yield was 36.11 kg ± 15.6. Of the 35 cows selected, 13 did not have any metritis during the study period, while 22 were diagnosed at least once with metritis (score 2, or higher), occurring on average at 12 DIM (12.02 DIM ± 4.72). Among these, 2 had retained fetal membranes and were kept for data analysis. None of the selected animals had hyperketonemia, mastitis, or hypocalcemia. The proportion of metritis events among primiparous was 20%, while the proportion of metritis events among multiparous was 23%. Based on the changes of metritis score between two consecutive evaluations, 239 health events were created, and of those, 188 were in the non-metritis event class, while 51 were in the metritis event class, resulting in an unbalanced dataset. Across all modeling steps, time windows,

and time lags, best performance in terms of F_1 score, Se, and PPV was obtained when the 20% threshold was used.

4.4.1 Nowcasting Framework: Individual Devices

When comparing the different levels of sensor data aggregation, there was a trend in model performance across the different aggregations. During modeling step 1, overall performance decreased as time windows became narrower, this is, the number of features used to fit the model decreased. However, the width of the time window at which performance started to drop varied depending on the level of aggregation of the time series data. Figure 4.1 shows that when sensor data were aggregated using 12 hour time window, performance dropped starting at -4 time steps before event, while performance for sensor data aggregated using 6 hour time window dropped starting at -5 and -4 time steps before the event for CowManager and TrackaCow, respectively. Similarly, when sensor data were aggregated using 3 hour time window, performance started to drop at -6 time steps before the event. Therefore, the smaller the level of data aggregation, the longer F_1 score value remained high before starting to drop. F_1 score oscillations were more evident for 6 hour *versus* 3 hour time windows. Similar trend could be observed for modeling step 2, where for all levels of data aggregation, classifier performance decreased with narrower time windows. However, at modeling step 2, performance at narrower time lags did not decrease as acutely as during modeling step 1 (Figure 4.1). Less clear trends could be observed for classifier performance at modeling step 3. Overall, adding a milk yield-related variable during modeling improved the F_1 score at -1 time steps before the event with respect to modeling step 2. However, the improvement in F_1 score was

greater at wider time windows: between -1.5 and -3 days before the event with sensor data aggregated using 12 hour time windows, and between -1.75 and -12 days before the event with sensor data collected with CowManager device and 6 hour time windows. Improvement of F_1 score during modeling step 3 was less consistent with sensor data registered by TrackaCow device and 3 hour time window, being constantly higher between -2.5 and -3 days before the event (Figure 4.1).

Table 4.1 shows the most relevant performance metrics at the different modeling steps when CowManager data were aggregated using a 12 hour time window. Complete list of performance metrics from all the models can be found in a data repository (Vidal et al.). At modeling step 1, the behaviors with greatest F_1 scores were eating and ruminating, with Se values that ranged between 85.4 – 92.5% and 61.9 – 82.9%, respectively, and with PPV values that ranged between 92.1 – 100% and 63.4 – 87.2%, respectively (Table 4.1). At modeling step 2, the addition of multiple behaviors only improved performance at narrower time windows (models 4 through 1; Table 4.1). At modeling step 3, the milk yield-related variable that yielded greatest F_1 score was the variance of the daily milk yield for the previous 3 days before the metritis event. Nevertheless, for those models where adding this feature improved performance it was at wider time lags, with an F_1 score that increased by 2.5%, 4.1%, and 1.5% for models 4, 5, and 6, respectively (Table 4.1). Adding the milk yield variance to models 1, 2, and 3 lowered the F_1 score, a result that was driven by a small decrease on PPV compared to a marginal improvement of Se (Table 4.1). Other milk yield transformations were explored, with some of these increasing the Se but lowering the PPV, or vice versa.

When sensor data measured with CowManager device were aggregated using a 6 hour time window, behaviors with best performance in terms of F_1 score at modeling step 1 were for those associated with different levels of activity (not active, active, and high activity behaviors; Table 4.2). At modeling step 2, best performances were also found in models where behaviors corresponding to different levels of activity were combined (models 12 through 18), although adding more features from multiple behaviors did not always improve the F_1 score obtained during modeling step 1, especially for models 12 through 18 (Table 4.2). In those cases where an improvement in the F_1 score was observed between steps 1 and 2, it was driven by an increment in both, Se and PPV. With respect to modeling step 1, at modeling step 2 F_1 score increased by a range of 8.6% - 47.7% for models 7 through 10, and by 2.8% - 54% for models 11, 14, and 16. At modeling step 3, best milk-related features were either the daily milk yield, or the daily milk yield variance, although 100% F_1 score was achieved in models 14 and 11 using the milk yield variance as model input. At modeling step 3, there was a tradeoff between Se and PPV, with greater F_1 score than those obtained at step 2 that were driven by an improvement in the Se, outweighing a poorer performance in terms of PPV when compared with modeling step 2 (Table 4.2).

When sensor data measured with TrackaCow device were aggregated using a 6 hour time window, behaviors with best performance in terms of F_1 score at modeling step 1 were for lying time, followed by number of steps (Table 4.3). During modeling step 2, the best combinations of behaviors were for lying time and number of steps, with the exception of model 19, where features from all behaviors measured with TrackaCow had to be added in order to improve the F_1 score (Table 4.3). Overall, greater improvement in F_1 score was observed at narrower time windows driven by both, Se and PPV: F_1 score at modeling step 2 improved by 6% to 42% for

models 22 through 19. Greatest F_1 score was obtained at step 3 by adding the milk yield variance (models 27, 28, and 29; Table 4.3). For models 23 through 26, adding either daily milk yield or milk yield variance did either improved Se or PPV, but not both, resulting in a lower F_1 score that the one estimated during modeling step 2. For the models where daily milk yield improved performance, F1 score increased by 0.44 – 5.2% (models 19 through 22, and 27 through 29), and from 0.3% to 4% (models 36, and 39 through 53; Table 4.3).

When sensor data measured with TrackaCow device were aggregated using a 3 hour time window, behaviors with best performance in terms of F_1 score at modeling step 1 were for lying time and number of steps (Table 4.4). At modeling step 2, the best combinations were either lying and lying bout, or lying and steps (Table 4.4). Nevertheless, no improvement of the F1 score at modeling step 1 was found by adding multiple behaviors at model step 2, with the exception of narrower time lags (models 31 through 36). Therefore, and except for narrow time lags, simpler models from modeling step 1 were selected as best models before attempting to add milk yield-related variables in modeling step 3. Greatest F_1 scores were obtained when features from milk variables were added to the best models from modeling steps 1 and 2, being most of these models that included features from one behavior (Table 4.4). Among the milk yield transformation, the raw daily milk yield in the previous 3 days before the event was the best feature, but the improvement was marginal: from 0.3% - 4% for models 31 through 36, and 39 through 53. In almost all cases, the increment in the F_1 score was driven by an increment in Se, while 100% PPV was achieved in most cases during modeling step 1.

4.4.2 Nowcasting Framework: Combination of Devices

The only time window at which performance resulting from combining both, CowManager and TrackaCow devices, could be explored was the 6 hour time window. Overall, TrackaCow had greater F_1 scores across all time windows compared to CowManager (Figure 4.2), and those were obtained with fewer features. Furthermore, at narrower time windows, the drop in performance for CowManager was sharper (from 97.5% at time -5 to 89% at time -1; Figure 4.2) than the drop for TrackaCow (from 97.7% at time -5 to 94.5% at time -1; Figure 4.2). In any case the combination of the best models from each device improved performance, except at time before event -1, when performance was comparable to that one obtained with TrackaCow: 94.4% and 94.7% at time before event -1 for TrackaCow and the combination of TrackaCow with CowManager, respectively. The drop in performance was driven by a drop in Se, even though PPV either increased to 100%, or stayed at 100% when compared to individual device performance. Among milk-related variables, milk yield variance had the best performance compared with the others. However, in any case did daily milk yield or its transformations improved model performance compared to using behavioral variables alone, with the only exception of time step -1, when F_1 score reached 100% by adding either mean milk yield, milk yield variance, or milk yield slope into the model (Table 4.5).

4.4.3 Forecasting Framework

Finally, we also estimated model performance under a forecasting framework. Best performance for CowManager was found 2 days forward before an event with the combination of active and eating behaviors using a 12 hour time window (Se = 92.68%, PPV = 100%, F_1 score =

96.2 at modeling step 2; Table 4.6), representing an increment by 35.7% with respect to modeling step 1. In contrast, ruminating behavior yielded best performance 3 days forward before an event (Se = 92.5%, PPV = 100%, F_1 score = 96.1 at modeling step 1; Table 4.6). At modeling step 3, the addition of milk yield only improved F_1 score with the 6 hour time window, improving by 0.2% and 1.4% for the 2 and 3 days forward predictions, respectively.

Best performance for TrackaCow was found 2 days forward and modeling step 1 with a time window of either 6 hours (Se = 95.35%, PPV = 100%, F_1 score = 97.62%; Table 4.7), or 3 hours (Se = 95.24%, PPV = 100%, F_1 score = 97.56%; Table 4.7). For the 3 days forward predictions, the best model was for lying behavior (Se = 95%, PPV = 97.44%, F_1 score = 96.2% and modeling step 3; Table 4.7). Improvement in performance was only observed 3 days forward and 6 h time window, with F_1 score improving by 2.7% between modeling step 1 and 2, and by 1.3% between modeling step 2 and 3, driven by an increment in PPV that outweighed the decrease in Se between modeling step 2 and 3.

4.5 Discussion

In this study, we assessed the performance in terms of Se, PPV, and F_1 score of a RF classifier for the prediction of metritis events, using a total of ten animal behaviors measured by two PDFTs and aggregated at two different time windows under multiple time steps. Furthermore, we selected our best models using three distinct modeling steps based on model complexity, exploring changes in behavioral variables performance when milk yield variables were added into the models. Additionally, we explored RF performance when best selected models from both PDFTs were combined into one model. Finally, all the different combinations

of sensor data aggregation and time windows were explored under nowcasting and forecasting frameworks. Our results address the current need for a systematic approach to the study of appropriate level of sensor data aggregation and number of time steps included as features for metritis prediction using different animal behaviors registered by PDFTs, and their combinations. In our study, we had an unbalanced dataset and therefore, we used a rank-based method approach followed by the comparison in performance under different thresholds (Vidal et al., Chapter 1 and 2). Many software algorithm implementations use by default the 50% class probability as threshold to classify observations into cases and non-cases, however, changing the default has been proposed as an strategy to improve classifier performance (Ouellet et al., 2016; Steensels et al., 2016). Our results showed that best results in terms of Se, PPV and F_1 score were obtained at the 20% threshold across all devices, time windows, and number of time steps. This threshold was closer to the prevalence of metritis events in our sample than the 30% and 40% thresholds. We suggest that in those case where animals are being categorized according to one or several underlying continuous traits, different thresholds should be tested, since not only PPV and NPV change with prevalence but also do Se and Sp (Brenner and Gefeller, 1997).

4.5.1 Nowcasting Framework: Individual Devices

Our results under nowcasting framework showed that, overall F_1 score across multiple levels of sensor data aggregation was lower at narrower time windows when only the features corresponding to one behavior were used (modeling step 1). For those narrower time window, F_1 score improved when features from more than one behavior were added to the model (modeling step 2), obtaining only a marginal improvement when milk yield was added to the

model (modeling step 3). The changes in RF performance observed in our study were directly influenced by the increase in the number of features between modeling steps 1 and 2 under the nowcasting framework. This effect can be attributed to the bias-variance trade-off, this is, the relationship with the expected prediction error and the mean squared error of our predictions. More generally, as the model complexity increases (higher number of features), the variance tends to increase and the squared bias (the amount by which the average of our estimate differs from the true mean) tends to decrease (Hastie et al., 2009). In our study, at narrower time windows for modeling step 1, where fewer features were used to fit the model, models may have been too simplistic and underfitting may have occurred. In contrast, when for the same time windows more features were added during modeling steps 2 and 3, resulting in an increase in model complexity with the associated risk of overfitting. Fine tuning the complexity of prediction models will need to be explored in those cases where classification algorithms will be implemented either as validation studies or as clinical trials in commercial farms.

In our study we found that classifier performance changed across the different levels of sensor data aggregation. Under the nowcasting framework and 12 h time window, several behaviors measured with CowManager device were useful when classifying events at modeling step 1, being eating preferred for wider time windows while ruminating was preferred for narrow time windows. In contrast, features from almost all behaviors measured with CowManager were needed in order to increase model complexity so an improvement in performance at narrower time windows could be achieved. For those behaviors measured with TrackaCow under a nowcasting framework, features corresponding to behaviors lying, steps, or to the combination of both, showed best classification performance at both, 6 h and 3 h time

windows. When the former time window was used, greater improvement in F_1 score was observed between modeling step 1 and 2 with fewer time steps included in the model. The improvement in F_1 score was driven by both, Se and PPV. The study of the impact of different time windows has been mainly investigated in sensor research, where animal behaviors are predicted based on sensor signals (Walton et al., 2018), but it has been poorly studied in research where changes in animal behavior are being used to predict different health states (Carslake et al., 2021). Comparison of our findings with previous studies was not completely possible, as performance at different levels of data aggregation and time windows have not been systematically explored for disease detection (Carslake et al., 2021). Previous studies have investigated the performance of a RF classifier to classify estrus events using all behaviors measured with CowManager device with 12 h time windows (Dolecheck et al., 2015), finding a wider range of values for Se (47.82 – 100%) compared to the models that used 12 h time windows in our study (90 – 95.24%, models 1 through 6). The study by Dolecheck et al. (2015) did not report PPV, but differences in Se could be due to the fact that their dataset was small, unbalanced, and features from all behaviors were combined into one model, an strategy we have proven in this study as not necessarily beneficial for classifier performance (Vidal et al., Chapter 1 and 2). Steensels et al. (2016) used a decision-tree model to detect a combination of ketosis, metritis, or both, during postpartum. In their model, they combined rumination, activity, milk yield, milk slope, and body weight change since parturition (Steensels et al., 2016). Using different time windows than those used in this study (2 h for activity and 24 h for rumination), reported Se was 86% and PPV was 88%. These values are lower than the mean values (± SD) we found in our study for rumination and activity behaviors using either 12 h or 6 h time windows

and for multiple time window widths at modeling step 3 (for 12 h time window: mean Se 94.66% ± 3.76; mean PPV 96.49% ± 4.48; for 6 h time window: mean Se 97.64 ± 3.43; mean PPV 94.88 ± 4.66). Differences between both studies could be due to the fact that both, ketosis and metritis were combined, making the assumption that the patterns in the data for ketosis and metritis may be equivalent while that may not be true. Furthermore, Steensels et al. (2016) used a 24 h time window for the behavior ruminating, a level of aggregation for sensor data we found might not be the preferred one (Vidal et al., Chapter 1).

Among all the features related with milk yield explored in modeling step 3, milk yield variance and daily milk yield from the previous 3 days before an event had best performance compared to milk yield mean and milk yield slope. Among all the models and time windows, milk yield variance was preferred for CowManager, while daily milk yield was preferred for TrackaCow in terms of performance. Nevertheless, adding milk yield variables to the best selected models from modeling steps 1 and 2 did not always improve performance, and sometimes improvement was only observed at wider time windows, or the improvement was marginal. Generally, at modeling step 3, an improvement in F_1 score was driven by an increase in Se that outweighed a decrease in PPV. Our findings are different from those found by others: a study by Steensels et al. (2016) included milk yield slope in their models for ketosis and metritis, while in our study, milk yield slope did not perform well. A more recent study has found that milk yield perturbations last between 5 and 207 days (mean 19.8 days ± 20.7; Adriaens et al., 2021). In light of these findings, we hypothesized that our 3 day time window for milk yield-related variables may be too small. Future studies should include wider time windows (at least 5 days before

event) and different milk yield transformations should be explored, as these may be characteristic of different diseases (Adriaens et al., 2021).

Additionally, prioritization of one performance metric over the others is problem-specific: in some cases, such as estrus detection, capturing all true positives at the expense of a higher false positive rate would be preferred due to the cost associated with missed events, while in other cases such as illness detection or calving events, high number of false positives will cause financial losses due to unnecessary treatment (Borchers et al., 2017). The advantage of using F_1 score is that the weights for the Se and PPV can be modified to adapt to different management scenarios (Vidal et al., Chapter 1 and 2).

4.5.2 Nowcasting Framework: Combination of Devices

In this study, we were able to explore the RF classifier performance using behaviors registered by both, CowManager and TrackaCow devices, and 6 h time window for the aggregation of the sensor data. Overall, TrackaCow had greater F_1 scores at all time windows compared to CowManager, and these were obtained with simpler models. Our findings are in agreement with those found by Tsai (2017) and Lee (2018) during their primary analysis of the dataset used in our study. Despite the fact that different time windows and classifiers were used to identify cows with metritis, Tsai (2017) found that TrackaCow had a Se of 75% while the Se for CowManager was 57%. Similarly, Lee (2018) found that TrackaCow had a Se of 77% while CowManager had a Se of 57%. Although the combination of CowManager and TrackaCow had not been studied, primary data analysis showed that when behavioral data registered by TrackaCow was combined with the behavioral data registered by other devices other than

CowManager, performance of TrackaCow in terms of Se and Sp worsen (Lee, 2018). Similar results were found for the combination of different devices registering rumination, steps, lying time, and lying bouts, where PPV decreased compared with performances for each device alone (Borchers et al., 2017). Performance resulting from the combination of multiple devices has also been studied to classify reproductive events. A previous study combined rumination time, lying time, and lying bouts using 6 h time windows with logistic regression as classifier for calving events, reporting a Se that ranged from 42% to 86%, and a PPV between 10% and 23% (Ouellet et al., 2016). In our study, none of the best selected models contained the same combination of behaviors, and instead, different levels of activity, lying time, and steps yielded best results, with greater mean (± SD) Se and PPV (86.5% ± 5.4 and 98.5% ± 3.6, respectively) that those reported by Ouellet et al. (2016).

4.5.3 Forecasting Framework

Forecasting frameworks are studied in the prediction of infectious diseases (Rashid, 2003; Thompson and Brooks-Pollock, 2019). To our knowledge, these have not been compared with nowcasting frameworks for a given dataset in PDFT literature. For this study, we explored the differences in performance under a forecasting framework, using 2 and 3 days forward forecasts. For the behaviors measured by CowManager, improvement in performance was only observed for the combination of 12 h time window and 2 days forward, while for those behaviors measured by TrackaCow, improvement was only observed for the combination of 6 h time window and 3 days forward. Even though differences between the best models for CowManager were not very different in terms of Se and PPV, the advantage of identifying a sick

cow 3 days earlier may outweighs small differences in Se and PPV. Similarly, the selection of the number of days forward for the forecast may change depending on if we want to prioritize a high PPV over a lower Se.

Results under nowcasting and forecasting frameworks are not directly comparable, because the features included in the model to make the predictions are different. Still, it is worth noting that under a nowcasting framework, a similar performance in terms of Se was obtained when all sensor data from the previous 2 days registered by CowManager was used to fit the model (model 4), with the difference that under a nowcasting framework, producers won't be able to implement earlier medical interventions. Therefore, for the same level of performance, we could have a Se of 92.68% and PPV of 100% by combining sensor data from 2 days before the event for behaviors active and eating.

In this study, the performance of our best selected forecasting models was greater than 90% Se and PPV, which is higher than those reported in previous studies. Ouellet et al., (2016) found that lying time could predict calving 1 day forward with a Se between 47% and 65%, and a PPV between 26% and 39%. Slightly higher values were reported by Borchers et al., (2017) for calving prediction using an 8 hour forward forecast, with a Se between 65.5% and 72.4%, and a PPV between 67.9% and 77.8%. Straightforward comparison with other studies was not possible due to methodological differences, and discrepancies between our finding and those by others could be due to the classifier used, threshold chosen, and time windows used to aggregate the data, as we know these impact performance (Vidal et al., Chapter 1 and 2).

4.5.4 Future Directions

Given the high volume of data generated by sensor devices and the relative novelty of the precision dairy farming field, finding multiple studies where same methodology and approach were used was challenging. Further studies are needed to identify the behavioral variables that may improve classification performance under different levels of sensor data aggregation and time windows. Furthermore, we showed how the combination of data registered by multiple devices was not beneficial in this particular dataset. Therefore, we also need a better understanding of return on investment of combining and connecting multiple sensor devices, and the technological challenges of when integrating diverse sources of data (Rutten et al., 2013). Lastly, forecasting frameworks can be advantageous over nowcasting frameworks since earlier interventions can be implemented. Nevertheless, further studies are needed to understand the cost of premature interventions, and the cost of technical support (Borchers, 2015). Limitations remain with current prediction models regarding how to deal with cases of more than one illness, and how to predict one illness without excluding others from the analyses.

4.6 Conclusions

This study presents a new methodology to study the optimal number and types of behaviors measured with PDFTs to predict metritis events. We have developed a framework for model building, allowing a better understanding of the interactions between model complexity, level of sensor data aggregation, and time windows, and the value of adding milk yield-related variables for improved performance. This study has shown that good classification performance

can be achieved with simple models involving one or two behavioral variables. The second major finding was that the combination of TrackaCow with CowManager and 6 h time windows does not improve performance compared with using TrackaCow only. Lastly, we found that when CowManager data were aggregated using 12 hour time windows, forecasting models can allow prediction of metritis 1 day earlier than nowcasting models.

4.7 Acknowledgements

The authors would like to thank the University of Kentucky Coldstream dairy staff, and to all the students who helped with fresh cow exam and data collection. We would also like to thank Jeffrey Bewley for facilitating data sharing. The work was partially supported by NSF award IIS-BigData-AI-1838207. JS is partially supported by NSF DMS 1712996.

Table 4. 1: Metrics (%) used for model building under nowcasting framework at each one of the modeling steps for behaviors measured with an ear-attached 3-axis accelerometer (CowManager, Agis Autimatisering, Harmelen, Netherlands), with sensor data aggregated using 12 hour time windows and using random forest to classify metritis events.

Model ID		**6**	**5**	**4**	**3**	**2**	**1**
Time Lag *k*		-6	-5	-4	-3	-2	-1
Step 1[3]	Behaviors[1]	E	E	E	R	R	R
	Nfeatures[2]	6	5	4	3	2	1
	Se	90	92.5	85.4	82.9	76.2	61.9
	PPV	100	100	92.1	87.2	80	63.4
	F_1 score	94.7	96.1	88.6	85	78.1	62.7
Step 2[3]	Behaviors[1]	E+HA	E+HA	R+E+NA	R+NA	R+E+NA	R+A+HA+NA
	Nfeatures[2]	12	10	12	6	6	4
	Se	90	92.5	92.7	95.1	95.2	92.9
	PPV	100	100	100	100	100	95.1
	F_1 score	94.7	96.1	96.2	97.5	97.6	94
Step 3[3]	Behaviors[1] and Milk Vars.[4]	E+MV	E+MV	R+E+NA+MV	R+NA+MV	R+E+NA+MV	R+A+HA+NA+MV
	Nfeatures[2]	7	6	13	7	7	5
	Se	92.6	100	96.4	89.3	96.6	93.1
	PPV	100	100	100	89.3	96.6	93.1
	F_1 score	96.2	100	98.2	89.3	96.6	93.1

[1] Behaviors: R: ruminating; E: eating; NA: not active; A: active; HA: high activity.

[2] Number of Features: model inputs were $\left(\bar{x}_{ij,t-1}, \bar{x}_{ij,t-2}, \ldots, \bar{x}_{ij,t-k}\right)$, where $\bar{x}$ was the mean of the hourly sensor values for behavior i and $j = 12\ h$, being the diagnosis assigned at 6:00 h on each one of the days when uterine discharge was evaluated, $i \in$ {ruminating, eating, not active, active, high activity}, and number of time steps before the event $k = 1, 2, \ldots, l$, where l was the number of time steps included as features within a 3 day period before a given metritis event.

[3] Modeling steps: Step 1: only features corresponding to one of the 5 behaviors measured with the device were used to fit the model; Step 2: features corresponding to multiple behaviors measured with a single device were used to fit the model; Step 3: features from milk yield-related variables were added to best model selected from steps 1 and 2.

[4] Milk Variables: MV: milk yield variance computed as the variance for the daily milk yield (kg) for the last 3 days prior to an event.

Table 4. 2: Metrics (%) used for model building under nowcasting framework at each one of the modeling steps for behaviors measured with an ear-attached 3-axis accelerometer (CowManager, Agis Autimatisering, Harmelen, Netherlands), with sensor data aggregated using 6 hour time windows and using random forest to classify metritis events.

Model ID		18	17	16	15	14	13	12	11	10	9	8	7
Time Lag *k*		-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1
Step 1[3]	Behaviors[1]	HA	NA	HA	A	A	HA	NA	NA	NA	E	NA	NA
	Nfeatures[2]	12	11	10	9	8	7	6	5	4	3	2	1
	Se	90	90	90	92.5	90	92.7	92.7	90.2	83	85.4	71.4	59.5
	PPV	100	100	97.3	100	97.3	100	100	94.9	87.1	87.5	75	61
	F_1 score	94.7	94.7	93.5	96.1	93.5	96.2	96.2	92.5	85	86.4	73.2	60.2
Step 2[3]	Behaviors[1]	HA+R	NA+A	HA+A	A+NA	A+HA+NA	HA+NA	NA+R	NA+E+HA+A	NA+R+E	E+HA+A	NA+R+E+A+HA	NA+R+E+HA+A
	Nfeatures[2]	24	22	20	18	24	14	12	20	12	9	10	5
	Se	90	90	92.5	92.5	92.5	92.5	92.7	95.1	95.1	92.7	90.5	88.1
	PPV	100	100	100	100	100	100	100	100	100	95	95	90.2
	F_1 score	94.7	94.7	96.1	96.1	96.1	96.2	96.2	97.5	97.5	93.8	92.7	89.2
Step 3[3]	Behaviors[1] and Milk Vars.[4]	HA+MY	NA+MY	HA+A+MV	A+MV	A+HA+NA+MV	HA+MV	NA+MY	NA+E+HA+A+MV	NA+R+E+MY	E+HA+A+MV	NA+R+E+A+HA+MS	NA+R+E+HA+A+MY
	Nfeatures[2]	15	14	21	10	25	8	9	21	15	10	11	8
	Se	100	100	96.3	96.3	100	92.9	100	100	100	96.4	89.7	100
	PPV	93.8	93.8	100	96.3	100	96.3	88.2	100	88.2	93.1	100	88.9
	F_1 score	96.8	96.8	98.1	96.3	100	94.6	93.8	100	93.8	94.7	94.6	94.1

[1] Behaviors: R: ruminating; E: eating; NA: not active; A: active; HA: high activity.

[2] Number of Features: model inputs were $\left(\bar{x}_{ij,t-1}, \bar{x}_{ij,t-2}, \ldots, \bar{x}_{ij,t-k}\right)$, where $\bar{x}$ was the mean of the hourly sensor values for behavior i and $j = 6\ h$, being the diagnosis assigned at 6:00 h on each one of the days when uterine discharge was evaluated, $i \in$ {ruminating, eating, not active, active, high activity}, and number of time steps before the event $k = 1, 2, \ldots, l$, where l was the number of time steps included as features within a 3 day period before a given metritis event.

[3] Modeling steps: Step 1: only features corresponding to one of the 5 behaviors measured with the device were used to fit the model; Step 2: features corresponding to multiple behaviors measured with a single device were used to fit the model; Step 3: features from milk yield-related variables were added to best model selected from steps 1 and 2.

[4] Milk Variables: MV: milk yield variance computed as the variance for the daily milk yield (kg) for the last 3 days prior to an event; MY: daily milk yield (kg) for each one of the 3 days before an event; MS: milk yield slope computed as the slope between maximum and minimum value for daily milk yield (kg) during the 3 days before an event.

Table 4. 3: Metrics (%) used for model building under nowcasting framework at each one of the modeling steps for behaviors measured with a leg-attached 3-axis accelerometer (TrackaCow, ENGS, Hampshire, UK), with sensor data aggregated using 6 hour time windows and random forest to classify metritis events. Models which performance metrics did not differ from other models have been omitted.

Model ID		**30**	**29**	**28**	**27**	**26**	**24**	**23**	**22**	**21**	**20**	**19**
Time Lag *k*		-12	-11	-10	-9	-8	-6	-5	-4	-3	-2	-1
Step 1[3]	Behaviors[1]	L	L	L	L	L	L	S	S	L	S	L
	Nfeatures[2]	12	11	10	9	8	6	5	4	3	2	1
	Se	95	92.7	95.1	95.2	95.4	95.4	95.6	88.9	86.4	73.3	65.2
	PPV	97.4	97.4	100	100	100	100	100	93	88.4	76.7	68.1
	F_1 score	96.2	95	97.5	97.6	97.6	97.6	97.7	90.9	87.4	75	66.7
Step 2[3]	Behaviors[1]	L+LB	L+S	L+S	L+LB	L+S	L+S	S+L	L+S	L+S	L+S	L+S+I+LB+IV
	Nfeatures[2]	24	22	20	18	16	12	10	8	6	4	5
	Se	97.5	95.1	92.7	95.2	95.4	95.4	95.5	95.5	95.5	95.5	89.5
	PPV	100	100	97.4	100	100	100	100	97.7	97.7	97.7	100
	F_1 score	98.7	97.5	95	97.6	97.6	97.6	97.7	96.7	96.6	96.6	94.4
Step 3[3]	Behaviors[1] and Milk Vars.[4]	L+LB+ MV	L+S+ MY	L+MY	L+MY	L+MY	L+MV+MY	S+MV	L+S+MY	L+S+MY	L+S+MY+ MV	L+S+I+LB+IV +MV
	Nfeatures[2]	25	25	11	12	11	10	6	11	9	8	6
	Se	96.2	100	100	100	100	100	93.3	100	100	100	91.7
	PPV	100	100	100	100	93.8	93.8	96.6	94.1	94.1	94.1	100
	F_1 score	98	100	100	100	96.8	96.8	94.9	97	97	97	95.7

[1] Behaviors: L: lying; LB: lying bouts; S: steps; I: intake; IV: intake visit.

[2] Number of Features: model inputs were $\left(\bar{x}_{ij,t-1}, \bar{x}_{ij,t-2}, \ldots, \bar{x}_{ij,t-k}\right)$, where $\bar{x}$ was the mean of the hourly sensor values for behavior i and $j = 6\ h$, being the diagnosis assigned at 6:00 h on each one of the days when uterine discharge was evaluated, $i \in$ {lying, lying bouts, steps, intake, intake visit}, and number of time steps before the event $k = 1, 2, \ldots, l$, where l was the number of time steps included as features within a 3 day period before a given metritis event.

[3] Modeling steps: Step 1: only features corresponding to one of the 5 behaviors measured with the device were used to fit the model; Step 2: features corresponding to multiple behaviors measured with a single device were used to fit the model; Step 3: features from milk yield-related variables were added to best model selected from steps 1 and 2.

[4] Milk Variables: MV: milk yield variance computed as the variance for the daily milk yield (kg) for the last 3 days prior to an event; MY: daily milk yield (kg) for each one of the 3 days before an event; MS: milk yield slope computed as the slope between maximum and minimum value for daily milk yield (kg) during the 3 days before an event.

Table 4. 4: Metrics(%) used for model building under nowcasting framework at each one of the modeling steps for behaviors measured with a leg-attached 3-axis accelerometer (TrackaCow, ENGS, Hampshire, UK), with sensor data aggregated using 3 hour time windows and random forest to classify metritis events. Models which performance metrics did not differ from other models have been omitted.

Model ID		**54**	**53**	**51**	**48**	**46**	**45**	**40**	**39**	**38**	**37**	**36**
Time Lag *k*		-24	-23	-21	-18	-16	-15	-10	-9	-8	-7	-6
Step 1[3]	Behaviors[1]	L	L	L	L	S	L	S	S	S	L	L
	Nfeatures[2]	24	23	21	18	16	15	10	9	8	7	6
	Se	92.5	95	92.7	92.7	93.2	93	93.3	95.6	95.6	95.5	93.2
	PPV	100	100	100	100	100	100	100	100	100	100	97.6
	F_1 score	96.1	97.4	96.2	96.3	96.5	100	96.6	97.7	97.7	97.7	95.4
Step 2[3]	Behaviors[1]	L+LB	L+LB	L+LB	L+S	S+L	L+S	S+L	S+L	S+L	L+S	L+S
	Nfeatures[2]	48	46	42	36	32	30	20	18	16	14	12
	Se	92.1	95	92.7	92.9	92.9	93	93.2	95.5	95.5	95.5	95.5
	PPV	100	100	100	100	100	100	100	100	100	100	100
	F_1 score	96.1	97.4	96.2	96.3	96.3	96.4	96.5	97.7	97.7	97.7	97.7
Step 3[3]	Behaviors[1] and Milk Vars.[4]	L+MV	L+MY	L+MY	L+MY	S+MY	L+MY	S+MY	S+MY	S+MY	L+MV	L+S+MY
	Nfeatures[2]	25	26	24	21	19	18	13	12	11	8	15
	Se	100	100	100	93.3	93.8	100	100	100	100	93.1	100
	PPV	100	100	100	100	100	100	100	100	94.1	96.4	100
	F_1 score	100	100	100	96.6	96.8	100	100	100	97	94.7	100

[1] Behaviors: L: lying; LB: lying bouts; S: steps; I: intake; IV: intake visit.

[2] Number of Features: model inputs were $\left(\bar{x}_{ij,t-1}, \bar{x}_{ij,t-2}, \ldots, \bar{x}_{ij,t-k}\right)$, where $\bar{x}$ was the mean of the hourly sensor values for behavior i and $j = 3\ h$, being the diagnosis assigned at 6:00 h on each one of the days when uterine discharge was evaluated, $i \in$ {lying, lying bouts, steps, intake, intake visit}, and number of time steps before the event $k = 1, 2, \ldots, l$, where l was the number of time steps included as features within a 3 day period before a given metritis event.

[3] Modeling steps: Step 1: only features corresponding to one of the 5 behaviors measured with the device were used to fit the model; Step 2: features corresponding to multiple behaviors measured with a single device were used to fit the model; Step 3: features from milk yield-related variables were added to best model selected from steps 1 and 2.

[4] Milk Variables: MV: milk yield variance computed as the variance for the daily milk yield (kg) for the last 3 days prior to an event; MY: daily milk yield (kg) for each one of the 3 days before an event; MS: milk yield slope computed as the slope between maximum and minimum value for daily milk yield (kg) during the 3 days before an event.

Table 4. 5: Metrics used for model comparison under nowcasting framework for the combination of the best models selected at modeling step 1 and 2 for both, ear-attached and leg-attached 3-axis accelerometers (CowManager and TrackaCow, respectively), and performance comparison when milk-related variables were added into the models. Random forest was used to classify metritis events. Only the first 5 time lags are shown.

Time Lag *k*	Metrics (%)	Both Devices	Milk Yield Variance[1]
-1	Se	100	100
	PPV	90	100
	F_1 score	94.7	100
-2	Se	77.8	73.3
	PPV	100	100
	F_1 score	87.5	84.6
-3	Se	82.4	71.4
	PPV	100	100
	F_1 score	90.3	83.3
-4	Se	82.4	71.4
	PPV	100	100
	F_1 score	90.3	83.3
-5	Se	87.5	76.9
	PPV	100	100
	F_1 score	93.3	87

[1] Milk Yield Variance was computed as the variance for the daily milk yield (kg) for the last 3 days prior to an event.

Table 4. 6: Metrics (%) used for model building under forecasting framework at each one of the modeling steps for behaviors measured with an ear-attached 3-axis accelerometer (CowManager, Agis Autimatisering, Harmelen, Netherlands), with sensor data aggregated using 12 and 6 h time windows and using random forest to classify metritis events.

Days Forward		2 days		3 days	
Time window		12H	6H	12H	6H
Step 1[3]	Behaviors[1]	A	E	R	A
	Nfeatures[2]	2	4	4	8
	Se	68.25	92.5	92.5	90
	PPV	73.68	100	100	100
	F_1 score	70.86	96.1	96.1	94.74
Step 2[3]	Behaviors[1]	A+E	E+R	R+NA+HA	A+R+NA
	Nfeatures[2]	4	8	12	24
	Se	92.68	92.5	92.5	90
	PPV	100	100	100	100
	F_1 score	96.2	96.1	96.1	94.74
Step 3[3]	Behaviors[1] and Milk Vars.[4]	A+E+MY	E+MY	R+MY	A+MY
	Nfeatures[2]	5	5	6	10
	Se	86.21	92.86	88.46	92.31
	PPV	92.59	100	95.83	100
	F_1 score	89.29	96.3	92	96

[1] Behaviors: R: ruminating; E: eating; NA: not active; A: active; HA: high activity.

[2] Number of Features: model inputs were $\left(\bar{x}_{ij,t-q}, \bar{x}_{ij,t-q-1}, \ldots, \bar{x}_{ij,t-l}\right)$, where $\bar{x}$ was the mean of the hourly sensor values for behavior i and $j = \{6\ h, 12\ h\}$, being the diagnosis assigned at 6:00 h on each one of the days when uterine discharge was evaluated, $i \in$ { ruminating, eating, not active, active, high activity }, and number of time steps before the event $k = q + 1, q + 2, \ldots, l$, where l was the number of time steps included as features within 24 hour period, and for $j = 6\ h$, $q \in \{4, 8\}$ for the 2 or 3 days forward, respectively, and for $j = 12\ h$, $q \in \{2, 4\}$ for the 2 or 3 days forward, respectively.

[3] Modeling steps: Step 1: only features corresponding to one of the 5 behaviors measured with the device were used to fit the model; Step 2: features corresponding to multiple behaviors measured with a single device were used to fit the model; Step 3: features from milk yield variables were added to best model from steps 1 and 2.

[4] Milk Variables: MY: daily milk yield (kg). Daily milk yield (kg) for the day -1 before event was not included in the forecasts.

Table 4. 7: Metrics (%) used for model building under forecasting framework at each one of the modeling steps for behaviors measured with a leg-attached 3-axis accelerometer (TrackaCow, ENGS, Hampshire, UK), with sensor data aggregated using 6 and 3 h time windows and using random forest to classify metritis events.

Days Forward		2 days		3 days	
Time window		6H	3H	6H	3H
Step 1[3]	Behaviors[1]	L	L	L	L
	Nfeatures[2]	4	8	8	16
	Se	95.35	95.24	92.5	95
	PPV	100	100	92.5	97.44
	F_1 score	97.62	97.56	92.5	96.2
Step 2[3]	Behaviors[1]	L+S	L+S	L+S	L+S+LB
	Nfeatures[2]	8	16	16	48
	Se	93.02	95.24	95	95
	PPV	97.56	100	95	97.44
	F_1 score	95.24	97.56	95	96.2
Step 3[3]	Behaviors[1] and Milk Vars.[4]	L+MY	L+MY	L+S+MY	L+S+LB+MY
	Nfeatures[2]	5	9	18	50
	Se	86.21	82.76	92.59	92
	PPV	96.15	96	100	100
	F_1 score	90.91	88.89	96.15	95.83

[1] Behaviors: L: lying; LB: lying bouts; S: steps; I: intake; IV: intake visit.

[2] Number of Features: model inputs were $\left(\bar{x}_{ij,t-q}, \bar{x}_{ij,t-q-1}, \ldots, \bar{x}_{ij,t-l}\right)$, where $\bar{x}$ was the mean of the hourly sensor values for behavior i and $j = \{3\ h, 6\ h\}$, being the diagnosis assigned at 6:00 h on each one of the days when uterine discharge was evaluated, $i \in$ { ruminating, eating, not active, active, high activity }, and number of time steps before the event $k = q + 1, q + 2, \ldots, l$, where l was the number of time steps included as features within 24 hour period, and for $j = 6\ h$, $q \in \{4, 8\}$ for the 2 or 3 days forward, respectively, and for $j = 3\ h$, $q \in \{8, 16\}$ for the 2 or 3 days forward, respectively.

[3] Modeling steps: Step 1: only features corresponding to one of the 5 behaviors measured with the device were used to fit the model; Step 2: features corresponding to multiple behaviors measured with a single device were used to fit the model; Step 3: features from milk yield variables were added to best model from steps 1 and 2.

[4] Milk Variables: MY: daily milk yield (kg). Daily milk yield (kg) for the day -1 before event is not included in the forecasts.

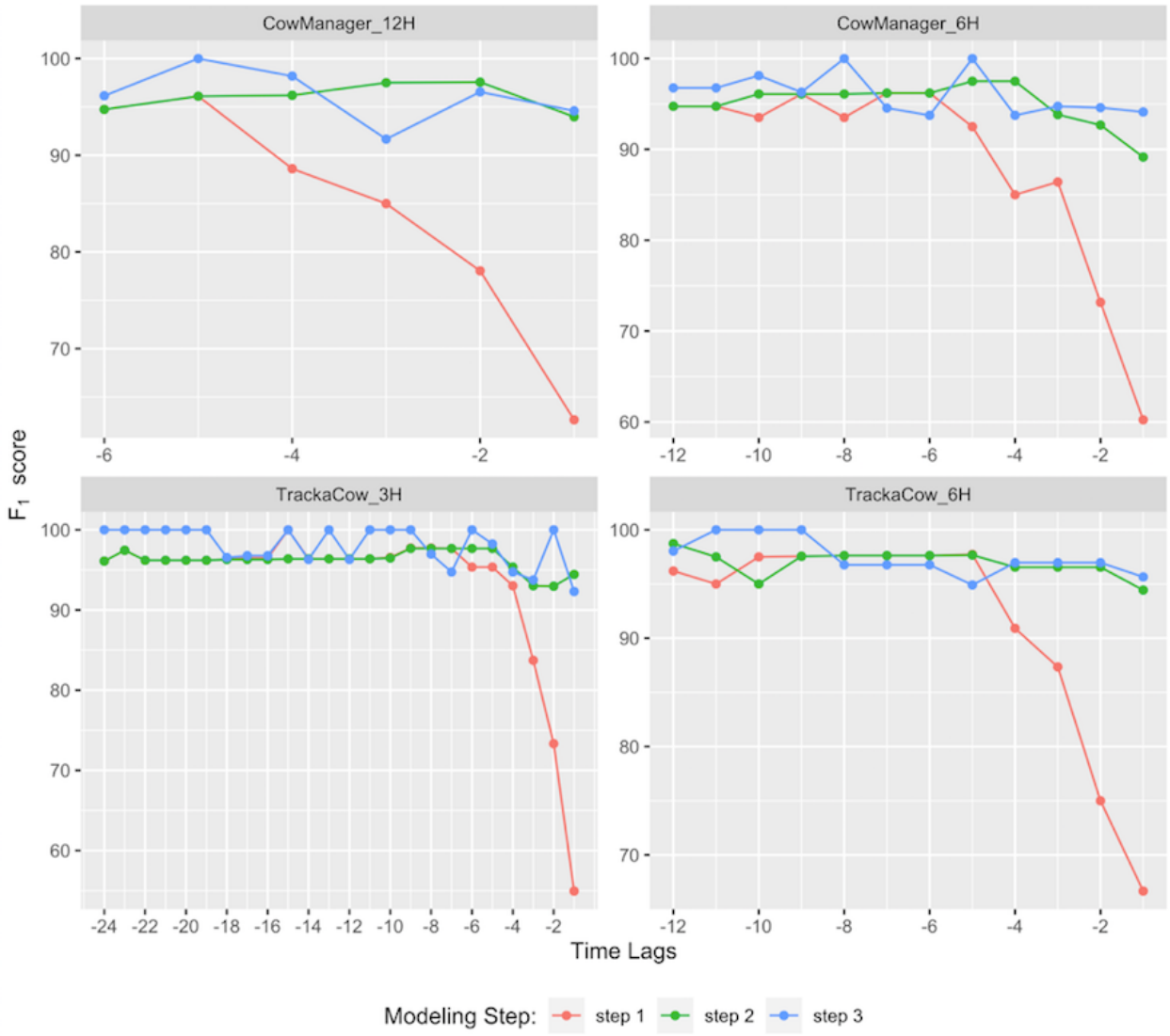

Figure 4. 1: Model performance comparison of the models with greatest F_1 score (%) at each modeling step, level of data aggregation (12, 6, and 3 hours), and number of time steps (time lags) before a metritis event.

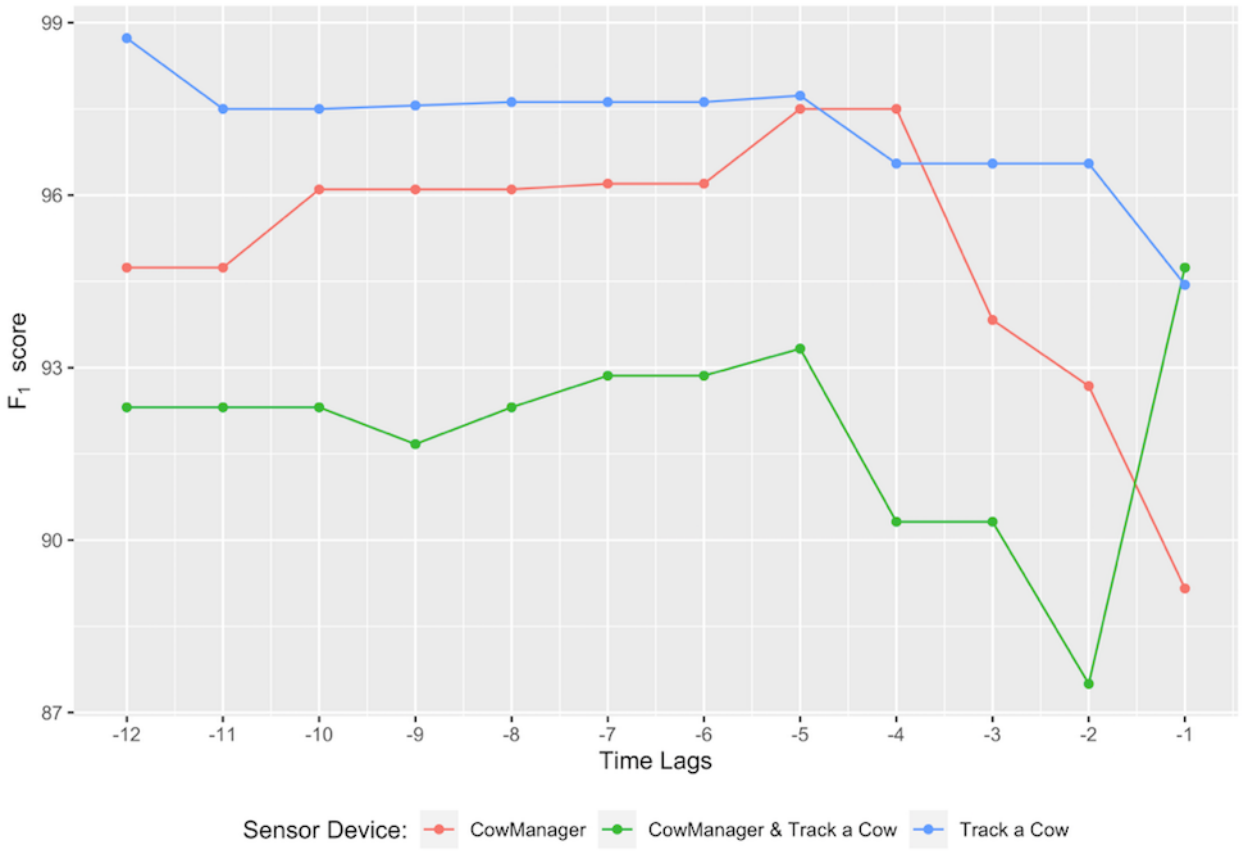

Figure 4. 2: Model performance comparison (F_1 score, %) before and after combining two devices, using random forest to classify metritis events and 6 hour time windows to aggregate sensor data. Models from each device to be combined were the best selected models from modeling step 1 and 2

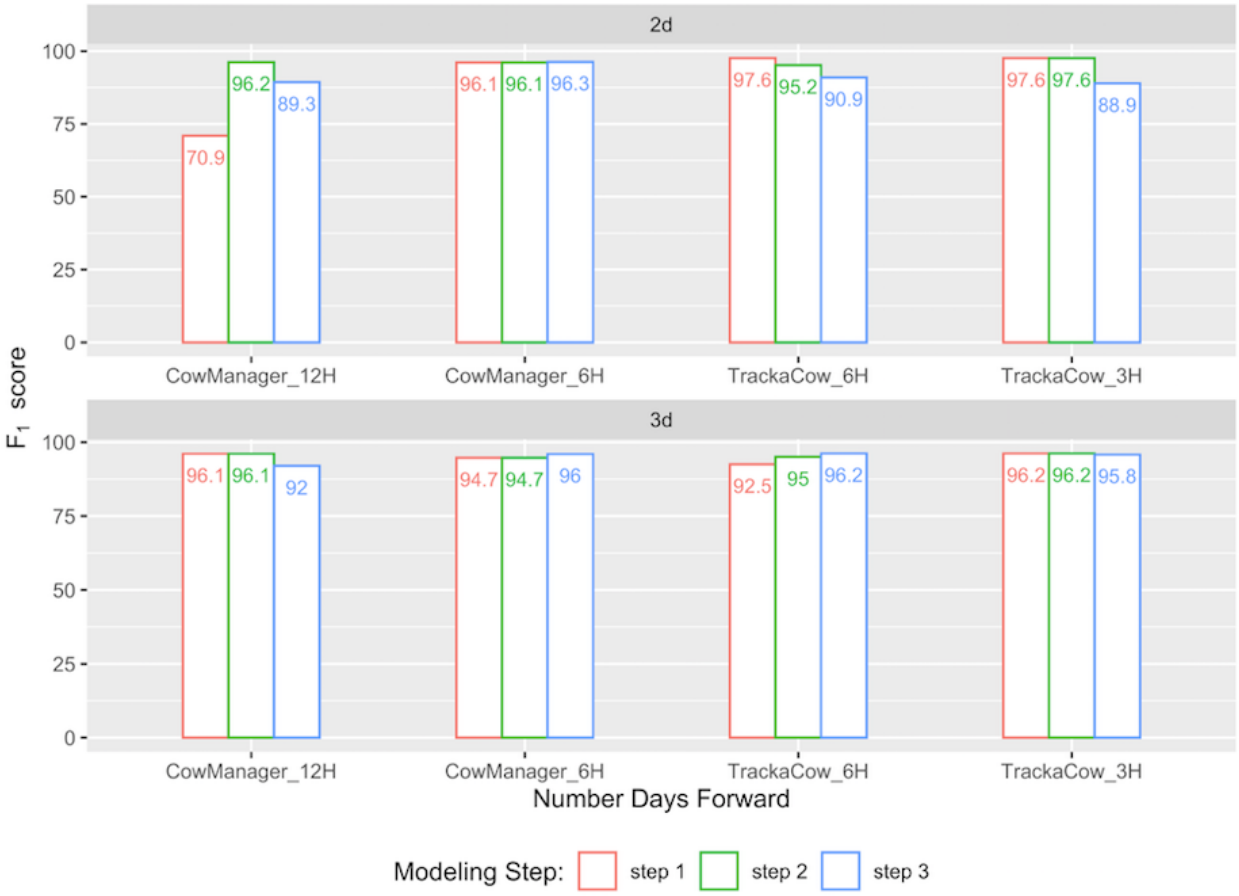

Figure 4. 3: Model performance (F_1 score, %) at each modeling step for each sensor device and time window under forecasting framework. Step 1: one model for each behavior was fitted and models were ranked from greatest to smallest F_1 score; step 2: models from step 1 were combined in a stepwise manner, starting with those with greatest F_1 score, Se, and PPV, in that order; step 3: milk yield-related variables were added independently to the best model selected from step 1 and 2, and the F_1 score of resulting model with greatest F_1 score, Se, and PPV was plotted.

.

5 Conclusions and Future Directions

My dissertation focused on the performance comparison of three machine learning supervised algorithms (*k*-nearest neighbors, random forest, and support vector machines) used for classification of metritis events in dairy cattle. To compare their performances, I used high frequency sensor data registered by two different devices: an ear-tag 3-axis accelerometer (chapter 2), and a leg-attached 3-axis accelerometer (chapter 3), and I combined data registered by these two devices with low frequency milk yield data (chapter 4). The whole research was conducted by controlling for other postpartum diseases such as hypocalcemia, hyperketonemia, and mastitis, while keeping temporal relationships between sensor measurements and metritis events.

Although PDFTs have been developing at a fast pace in recent years, very little is known about their prediction ability for postpartum diseases despite the high proportion of dairy cows suffering from these. Furthermore, when machine learning algorithms are used, most of the studies in PDFT literature lack a systematic workflow approach. For my research, I used a classical machine learning workflow that can be used to build other predictive models in the future for diseases affecting cow during postpartum such as hypocalcemia or hyperketonemia.

Relevant contributions were made in the field of PDFTs for metritis prediction by using a systematic approach to feature engineering and feature selection (chapter 2, 3, and 4). I studied the effect of within-same-day variability due to farm scheduled activities, and the interaction between number of days postpartum and farm scheduled activities (chapter 2 and 3). Even

though farm scheduled activities influenced animal behavior, better classification performance was achieved in those models where all sensor data were used, regardless of the time of the day. Based on our findings we further concluded that behavioral sensor data corresponding to the 3 days following parturition should be studied separately. Underlying mechanism such as those from social dynamics may add noise to the sensor data being used for disease prediction. Additionally, I compared classifier performance using multiple time windows (chapter 2, 3, and 4) and multiple lags (chapter 2, 3, and 4), contributing to our understanding of the impact that these have on classifier performance. I found that random forest yielded the best performance at predicting metritis events by using activity data, eating, and rumination time when time windows of 12 or 6 hours were used on sensor data (chapter 2). Similarly, random forest had the best performance at predicting those same metritis events by using lying time, number of lying bouts, and number of steps when time windows of 6 or 3 hours were used on sensor data (chapter 3).

There is a growing interest in using of ML algorithms to build predictive models for diseases postpartum using a combination of high frequency data registered by PDFTs and low frequency data that is being routinely collected such as production data. For my research I used a step wise approach to feature selection that included the study of classifier performance when features from multiple behaviors were added into the model, followed by the addition of milk yield-related variables. Additionally, I studied classifier performance when features from multiple behaviors registered by two different devices were combined into one model. I showed that good classification performance can be achieved with random forest and simple models where

features from one or two behavioral variables are included. My results were robust as similar findings were obtained across multiple time lags and time windows.

Another important aspect of my work is the contribution made in the evaluation of classifier performance to predict metritis events with unbalanced data as training sets (chapters 2, 3, and 4). Most studies found in the PDFT literature either do not address the challenges of using unbalanced datasets to train classifiers, or they artificially balance the dataset without addressing the limitations of implementing such practice in commercial farms. For my research, I used a rank-based approach to rank the observations based on their class probabilities (chapter 2, 3, and 4). As results, we obtained a higher sensitivity and positive predictive value than those reported by previous studies.

Lastly, another important contribution of my work was the comparison in performance for metritis prediction under nowcasting and forecasting frameworks (chapter 4). We found that similar results were yielded under both, nowcasting and forecasting frameworks using 12-hour time windows and 2 days forward before the metritis event. Consequently, the use of forecasting models could potentially identify cows at higher risk of developing metritis earlier than traditional diagnostic methods.

Despite our promising results regarding model performance across chapter 2, 3, and 4, model assessment on a test set remains a challenge due to limited amount of clinical data available. Similarly, model tunning when machine learning workflows are implemented on farms may cause important differences in performance compared with those seen here. This may be due to differences in feature importance, interactions with other concurrent diseases, different

selection of thresholds for classification, or different costs associated with the balance between positive predictive value and sensitivity.

6 Bibliography

Adriaens, I., I. van den Brulle, L. D'Anvers, J.M.E. Statham, K. Geerinckx, S. De Vliegher, S. Piepers, and B. Aernouts. 2021. Milk losses and dynamics during perturbations in dairy cows differ with parity and lactation stage. J. Dairy Sci. 104:405–418. doi:10.3168/jds.2020-19195.

Alpaydin, E. 2010. Introduction to Machine Learning. The MIT Press.

AlZahal, O., H. AlZahal, M.A. Steele, Van Schaik, I. Kyriazakis, T.F. Duffield, and B.W. McBride. 2011. The use of a radiotelemetric ruminal bolus to detect body temperature changes in lactating dairy cattle. J. Dairy Sci. 94:3568–3574. doi:10.3168/jds.2010-3944.

Alzahal, O., M.A. Steele, E. V. Valdes, and B.W. McBride. 2009. Technical note: The use of a telemetric system to continuously monitor ruminal temperature and to predict ruminal pH in cattle. J. Dairy Sci. 92:5697–5701. doi:10.3168/jds.2009-2220.

Andermann, P., S. Schlögl, U. Mäder, M. Luster, M. Lassmann, and C. Reiners. 2007. Intra-and interobserver variability of thyroid volume measurements in healthy adults by 2D versus 3D ultrasound. Nuklearmedizin 46:1–7.

Baird, G.D. 1982. Primary ketosis in the high-producing dairy cow: clinical and subclinical disorders, treatment, prevention, and outlook. J. Dairy Sci. 65:1–10.

Bar, D., and R. Soloman. 2010. Rumination Collars: What Can They Tell Us. First North Am. Conf. Precis. Dairy Manag. 2010 2. doi:10.1299/kikaic.79.4003.

Barragan, A.A., J.M. Piñeiro, G.M. Schuenemann, P.J. Rajala-Schultz, D.E. Sanders, J. Lakritz, and S. Bas. 2018. Assessment of daily activity patterns and biomarkers of pain, inflammation, and stress in lactating dairy cows diagnosed with clinical metritis. J. Dairy Sci. 101:8248–

8258. doi:10.3168/jds.2018-14510.

Beauchemin, K.A. 2018. Invited review: Current perspectives on eating and rumination activity in dairy cows. J. Dairy Sci. 101:4762–4784. doi:10.3168/jds.2017-13706.

Benzaquen, M.E., C.A. Risco, L.F. Archbald, P. Melendez, M.-J. Thatcher, and W.W. Thatcher. 2007. Rectal Temperature, Calving-Related Factors, and the Incidence of Puerperal Metritis in Postpartum Dairy Cows. J. Dairy Sci. 90:2804–2814.

Bergstra, J., and Y. Bengio. 2012. Random search for hyper-parameter optimization.. J. Mach. Learn. Res. 13.

Bewley, J. 2010. Precision dairy farming: advanced analysis solutions for future profitability. Pages 2–5 in Proceedings of the first North American conference on precision dairy management, Toronto, Canada.

Bewley, J.M., R.E. Boyce, J. Hockin, L. Munksgaard, S.D. Eicher, M.E. Einstein, and M.M. Schutz. 2010. Influence of milk yield, stage of lactation, and body condition on dairy cattle lying behaviour measured using an automated activity monitoring sensor. J. Dairy Res. 77:1–6. doi:10.1017/S0022029909990227.

Bewley, J.M., M.E. Einstein, M.W. Grott, and M.M. Schutz. 2008. Comparison of reticular and rectal core body temperatures in lactating dairy cows. J. Dairy Sci. 91:4661–4672.

Bikker, J.P., H. van Laar, P. Rump, J. Doorenbos, K. van Meurs, G.M. Griffioen, and J. Dijkstra. 2014. Technical note: Evaluation of an ear-attached movement sensor to record cow feeding behavior and activity. J. Dairy Sci. 97:2974–2979. doi:10.3168/jds.2013-7560.

Bishop, C. 2006. Pattern Recognition and Machine Learning. Springer, New York.

Borchers, M.R. 2015. An evaluation of precision farming technology adoption, perception,

effectiveness and use 110.

Borchers, M.R., Y.M. Chang, K.L. Proudfoot, B.A. Wadsworth, A.E. Stone, and J.M. Bewley. 2017. Machine-learning-based calving prediction from activity, lying, and ruminating behaviors in dairy cattle. J. Dairy Sci. 100:5664–5674. doi:10.3168/jds.2016-11526.

Borchers, M.R., Y.M. Chang, I.C. Tsai, B.A. Wadsworth, and J.M. Bewley. 2016. A validation of technologies monitoring dairy cow feeding, ruminating, and lying behaviors. J. Dairy Sci. 99:7458–7466. doi:10.3168/jds.2015-10843.

Breiman, L. 2001. Random forests. Mach. Learn. 45:5–32. doi:10.1023/A:1010933404324.

Brenner, H., and O. Gefeller. 1997. Variation of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence. Stat. Med. 16:981–991. doi:10.1002/(SICI)1097-0258(19970515)16:9<981::AID-SIM510>3.0.CO;2-N.

Büchel, S., and A. Sundrum. 2014. Short communication: Decrease in rumination time as an indicator of the onset of calving. J. Dairy Sci. 97:3120–3127. doi:10.3168/jds.2013-7613.

Burfeind, O., K. Schirmann, M.A.G. Von Keyserlingk, D.M. Veira, D.M. Weary, and W. Heuwieser. 2011. Evaluation of a system for monitoring rumination in heifers and calves. J. Dairy Sci. 94:426–430.

Bzdok, D., N. Altman, and M. Krzywinski. 2018. Statistics versus machine learning. Nat. Methods 15:233–234. doi:10.1038/nmeth.4642.

Calderon, D.F., and N.B. Cook. 2011. The effect of lameness on the resting behavior and metabolic status of dairy cattle during the transition period in a freestall-housed dairy herd. J. Dairy Sci. 94:2883–2894.

Caraviello, D.Z., K.A. Weigel, M. Craven, D. Gianola, N.B. Cook, K. V. Nordlund, P.M. Fricke, and

M.C. Wiltbank. 2006. Analysis of reproductive performance of lactating cows on large dairy farms using machine learning algorithms. J. Dairy Sci. 89:4703–4722. doi:10.3168/jds.S0022-0302(06)72521-8.

Carslake, C., J.A. Vázquez-Diosdado, and J. Kaler. 2021. Machine learning algorithms to classify and quantify multiple behaviours in dairy calves using a sensor–moving beyond classification in precision livestock. Sensors (Switzerland) 21:1–14. doi:10.3390/s21010088.

Chapinal, N., M. Carson, T.F. Duffield, M. Capel, S. Godden, M. Overton, J.E.P. Santos, and S.J. LeBlanc. 2011. The association of serum metabolites with clinical disease during the transition period. J. Dairy Sci. 94:4897–4903. doi:10.3168/jds.2010-4075.

Chapinal, N., A.M. De Passillé, and J. Rushen. 2009. Weight distribution and gait in dairy cattle are affected by milking and late pregnancy. J. Dairy Sci. 92:581–588. doi:10.3168/jds.2008-1533.

Chapinal, N., D.M. Veira, D.M. Weary, and M.A.G. Von Keyserlingk. 2007. Technical note: Validation of a system for monitoring individual feeding and drinking behavior and intake in group-housed cattle. J. Dairy Sci. 90:5732–5736. doi:10.3168/jds.2007-0331.

Chaplin, S., and L. Munksgaard. 2001. Evaluation of a simple method for assessment of rising behaviour in tethered dairy cows. Anim. Sci. 72:191–197. doi:10.1017/S1357729800055685.

Correa, M.T., H. Erb, and J. Scarlett. 1993. Path analysis for seven postpartum disorders of Holstein Cows. J. Dairy Sci. 76:1305–1312.

Dado, R.G., and M.S. Allen. 1993. Continuous computer acquisition of feed and water intakes, chewing, reticular motility, and ruminal pH of cattle. J. Dairy Sci. 76:1589–1600.

Dasarathy, B. 1991. Nearest Neighbor Pattern Classification Techniques. IEEE Comput. Soc. Press. Los Alamitos, CA.

Davis, J., and M. Goadrich. 2006. The relationship between precision-recall and ROC curves. ACM Int. Conf. Proceeding Ser. 148:233–240. doi:10.1145/1143844.1143874.

DeVries, T.J., and M.A.G. Von Keyserlingk. 2005. Time of feed delivery affects the feeding and lying patterns of dairy cows. J. Dairy Sci. 88:625–631. doi:10.3168/jds.S0022-0302(05)72726-0.

DeVries, T.J., M.A.G. Von Keyserlingk, D.M. Weary, and K.A. Beauchemin. 2003. Technical note: Validation of a system for monitoring feeding behavior of dairy cows. J. Dairy Sci. 86:3571–3574. doi:10.3168/jds.S0022-0302(03)73962-9.

Dittrich, I., M. Gertz, and J. Krieter. 2019. Alterations in sick dairy cows' daily behavioural patterns. Heliyon 5:e02902. doi:10.1016/j.heliyon.2019.e02902.

Dolecheck, K.A., W.J. Silvia, G. Heersche, Y.M. Chang, D.L. Ray, A.E. Stone, B.A. Wadsworth, and J.M. Bewley. 2015. Behavioral and physiological changes around estrus events identified using multiple automated monitoring technologies. J. Dairy Sci. 98:8723–8731. doi:10.3168/jds.2015-9645.

Dominiak, K.N., and A.R. Kristensen. 2017. Prioritizing alarms from sensor-based detection models in livestock production-a review on model performance and alarm reducing methods. Comput. Electron. Agric. 133:46–67.

Donovan, G.A., C.A. Risco, G.D. Temple, T.Q. Tran, and H.H. Van Horn. 2004. Influence of transition diets on occurrence of subclinical laminitis in Holstein dairy cows. J. Dairy Sci. 87:73–84.

Dórea, F.C., A.R.W. Elbers, P. Hendrikx, C. Enoe, C. Kirkeby, L. Hoinville, and A. Lindberg. 2016. Vector-borne disease surveillance in livestock populations: A critical review of literature recommendations and implemented surveillance (BTV-8) in five European countries. Prev. Vet. Med. 125:1–9. doi:10.1016/j.prevetmed.2016.01.005.

Dórea, F.C., B.J. McEwen, W.B. McNab, J. Sanchez, and C.W. Revie. 2013. Syndromic surveillance using veterinary laboratory data: Algorithm combination and customization of alerts. PLoS One 8. doi:10.1371/journal.pone.0082183.

Drackley, J.K. 1999. Biology of dairy cows during the transition period: The final frontier?. J. Dairy Sci. 82:2259–2273.

Drackley, J.K., T.R. Overton, and G.N. Douglas. 2001. Adaptations of glucose and long-chain fatty acid metabolism in liver of dairy cows during the periparturient period. J. Dairy Sci. 84:E100–E112.

Drillich, M., O. Beetz, A. Pfützner, M. Sabin, H.-J. Sabin, P. Kutzer, H. Nattermann, and W. Heuwieser. 2001. Evaluation of a systemic antibiotic treatment of toxic puerperal metritis in dairy cows. J. Dairy Sci. 84:2010–2017.

Drissler, M., M. Gaworski, C.B. Tucker, and D.M. Weary. 2005. Freestall maintenance: Effects on lying behavior of dairy cattle. J. Dairy Sci. 88:2381–2387.

Dubuc, J., T.F. Duffield, K.E. Leslie, J.S. Walton, and S.J. LeBlanc. 2011. Effects of postpartum uterine diseases on milk production and culling in dairy cows. J. Dairy Sci. 94:1339–1346. doi:10.3168/jds.2010-3758.

Dupuy, C., A. Bronner, E. Watson, L. Wuyckhuise-Sjouke, M. Reist, A. Fouillet, D. Calavas, P. Hendrikx, and J.-B. Perrin. 2013. Inventory of veterinary syndromic surveillance initiatives in

Europe (Triple-S project): current situation and perspectives. Prev. Vet. Med. 111:220–229.

Edwards, J.L., and P.R. Tozer. 2004. Using activity and milk yield as predictors of fresh cow disorders. J. Dairy Sci. 87:524–531. doi:10.3168/jds.S0022-0302(04)73192-6.

Espadamala, A., P. Pallarés, A. Lago, and N. Silva-del-Río. 2016. Fresh-cow handling practices and methods for identification of health disorders on 45 dairy farms in California. J. Dairy Sci. 99:9319–9333. doi:10.3168/jds.2016-11178.

Firk, R., E. Stamer, W. Junge, and J. Krieter. 2002. Automation of oestrus detection in dairy cows: A review. Livest. Prod. Sci. 75:219–232. doi:10.1016/S0301-6226(01)00323-2.

Fix, E., and J.L. Hodges. 1951. Discriminatory analysis - nonparametris discrimination: consistency properties.

Földi, J., M. Kulcsar, A. Pecsi, B. Huyghe, C. De Sa, J. Lohuis, P. Cox, and G. Huszenicza. 2006. Bacterial complications of postpartum uterine involution in cattle. Anim. Reprod. Sci. 96:265–281.

Fourichon, C., H. Seegers, N. Bareille, and F. Beaudeau. 1999. Effects of disease on milk production in the dairy cow: A review. Prev. Vet. Med. 41:1–35. doi:10.1016/S0167-5877(99)00035-5.

Fregonesi, J.A., C.B. Tucker, and D.M. Weary. 2007. Overstocking reduces lying time in dairy cows. J. Dairy Sci. 90:3349–3354.

Geishauser, T., K. Leslie, D. Kelton, and T. Duffield. 1998. Evaluation of Five Cowside Tests for Use with Milk to Detect Subclinical Ketosis in Dairy Cows. J. Dairy Sci. 81:438–443. doi:10.3168/jds.S0022-0302(98)75595-X.

Géron, A. 2017. Hands-on Machine Learning with Scikit-Learn and Tensorflow. First edit. N.

Tache, ed. O'Reilley Media, Inc.

Gilbert, R.O., S.T. Shin, C.L. Guard, and H.N. Erb. 1998. Incidence of endometritis and effects on reproductive performance of dairy cows. Theriogenology 1:251.

Giuliodori, M.J., R.P. Magnasco, D. Becu-Villalobos, I.M. Lacau-Mengido, C.A. Risco, and R.L. de la Sota. 2013. Metritis in dairy cows: Risk factors and reproductive performance. J. Dairy Sci. 96:3621–3631.

Goff, J.P. 2006. Major advances in our understanding of nutritional influences on bovine health. J. Dairy Sci. 89:1292–1301.

Gomez, A., and N.B. Cook. 2010. Time budgets of lactating dairy cattle in commercial freestall herds. J. Dairy Sci. 93:5772–5781. doi:10.3168/jds.2010-3436.

Gröhn, Y., H.N. Erb, C.E. McCulloch, and H.S. Saloniemi. 1990. Epidemiology of reproductive disorders in dairy cattle: associations among host characteristics, disease and production. Prev. Vet. Med. 8:25–39.

Hadley, G.L., C.A. Wolf, and S.B. Harsh. 2006. Dairy cattle culling patterns, explanations, and implications. J. Dairy Sci. 89:2286–2296.

Haimerl, P., and W. Heuwieser. 2014. Invited review: Antibiotic treatment of metritis in dairy cows: A systematic approach. J. Dairy Sci. 97:6649–6661.

Hamilton, A.W., C. Davison, C. Tachtatzis, I. Andonovic, C. Michie, H.J. Ferguson, L. Somerville, and N.N. Jonsson. 2019. Identification of the Rumination in Cattle Using Support Vector Machines with Motion-Sensitive Bolus Sensors. Sensors 19. doi:10.3390/s19051165.

Hammon, D., I.M. Evjen, T.R. Dhiman, J.P. Goff, and J.L. Walters. 2006. Neutrophil function and energy status in Holstein cows with uterine health disorders. Vet. Immunol. Immunopathol.

113:21–29.

Hansen, S.S., P. Nørgaard, C. Pedersen, R.J. Jørgensen, L.S.B. Mellau, and J.D. Enemark. 2003. The effect of subclinical hypocalcaemia induced by Na 2 EDTA on the feed intake and chewing activity of dairy cows. Vet. Res. Commun. 27:193–205.

Hart, B.L. 1988. Biological basis of the behavior of sick animals. Neurosci. Biobehav. Rev. 12:123–137. doi:10.1016/S0149-7634(88)80004-6.

Hastie, T., R. Tibshirani, and J. Friedman. 2009. The Elements of Statistical Learning. Data Mining, Inference, and Prediction. Second Edi. Springer, New York.

Herskin, M.S., L. Munksgaard, and J. Ladewig. 2004. Effects of acute stressors on nociception, adrenocortical responses and behavior of dairy cows. Physiol. Behav. 83:411–420. doi:10.1016/j.physbeh.2004.08.027.

Hogeveen, H., C. Kamphuis, W. Steeneveld, and H. Mollenhorst. 2010. Sensors and clinical mastitis-the quest for the perfect alert. Sensors 10:7991–8009. doi:10.3390/s100907991.

Humblet, M.F., H. Guyot, B. Boudry, F. Mbayahi, C. Hanzen, F. Rollin, and J.M. Godeau. 2006. Relationship between haptoglobin, serum amyloid A, and clinical status in a survey of dairy herds during a 6-month period. Vet. Clin. Pathol. 35:188–193. doi:10.1111/j.1939-165X.2006.tb00112.x.

Huzzey, J.M., D.M. Veira, D.M. Weary, and M.A.G. Von Keyserlingk. 2007a. Prepartum behavior and dry matter intake identify dairy cows at risk for metritis. J. Dairy Sci. 90:3220–3233. doi:10.3168/jds.2006-807.

Huzzey, J.M., D.M. Veira, D.M. Weary, and M.A.G. Von Keyserlingk. 2007b. Prepartum behavior and dry matter intake identify dairy cows at risk for metritis. J. Dairy Sci. 90:3220–3233.

doi:10.3168/jds.2006-807.

Ingvartsen, K.L., and J.B. Andersen. 2000. Integration of metabolism and intake regulation: a review focusing on periparturient animals. J. Dairy Sci. 83:1573–1597.

Ito, K., N. Chapinal, D.M. Weary, and M.A.G. Von Keyserlingk. 2014. Associations between herd-level factors and lying behavior of freestall-housed dairy cows. J. Dairy Sci. 97:2081–2089.

Ito, K., M.A.G. von Keyserlingk, S.J. LeBlanc, and D.M. Weary. 2010. Lying behavior as an indicator of lameness in dairy cows. J. Dairy Sci. 93:3553–3560. doi:10.3168/jds.2009-2951.

Iwersen, M., U. Falkenberg, R. Voigtsberger, D. Forderung, and W. Heuwieser. 2009. Evaluation of an electronic cowside test to detect subclinical ketosis in dairy cows. J. Dairy Sci. 92:2618–2624.

Janeway Jr, C.A., P. Travers, M. Walport, and M.J. Shlomchik. 2001. Infectious agents and how they cause disease. Garland Science.

Jawor, P.E., J.M. Huzzey, S.J. LeBlanc, and M.A.G. Von Keyserlingk. 2012. Associations of subclinical hypocalcemia at calving with milk yield, and feeding, drinking, and standing behaviors around parturition in Holstein cows. J. Dairy Sci. 95:1240–1248.

Kamphuis, C., B. DelaRue, C.R. Burke, and J. Jago. 2012. Field evaluation of 2 collar-mounted activity meters for detecting cows in estrus on a large pasture-grazed dairy farm. J. Dairy Sci. 95:3045–3056. doi:10.3168/jds.2011-4934.

Kamphuis, C., H. Mollenhorst, A. Feelders, D. Pietersma, and H. Hogeveen. 2010a. Decision-tree induction to detect clinical mastitis with automatic milking. Comput. Electron. Agric. 70:60–68. doi:10.1016/j.compag.2009.08.012.

Kamphuis, C., H. Mollenhorst, J.A.P. Heesterbeek, and H. Hogeveen. 2010b. Detection of Clinical

Mastitis with Sensor Data from Automatic Milking Systems Is Improved by Using Decision-Tree Induction.

Kaneene, J.B., and R. Miller. 1995. Risk factors for metritis in Michigan dairy cattle using herd- and cow-based modelling approaches. Prev. Vet. Med. 23:183–200.

Karp, H.J., and C. Petersson-Wolfe. 2010. Use of milk lactose concentration as an indicator of mastitis following the validation of a novel in-line milk analysis system designed to measure milk components. Pages 1–2 in The first North American conference on precision dairy management.

Kaufman, E.I., S.J. LeBlanc, B.W. McBride, T.F. Duffield, and T.J. DeVries. 2016. Association of rumination time with subclinical ketosis in transition dairy cows. J. Dairy Sci. 99:5604–5618. doi:10.3168/jds.2015-10509.

Lacroix, R., F. Salehi, X.Z. Yang, and K.M. Wade. 1997. Effects of data preprocessing on the performance of artificial neural networks for dairy yield prediction and cow culling classification. Trans. ASAE 40:839–846.

LeBlanc, S. 2010. Monitoring metabolic health of dairy cattle in the transition period introduction—metabolic challenges in peripartum dairy cows and their associations with reproduction. J. Reprod. Dev. Reprod. Dev 56:29–35.

LeBlanc, S.J., T.F. Duffield, K.E. Leslie, K.G. Bateman, G.P. Keefe, J.S. Walton, and W.H. Johnson. 2002. Defining and diagnosing postpartum clinical endometritis and its impact on reproductive performance in dairy cows. J. Dairy Sci. 85:2223–2236.

LeBlanc, S.J., T. Osawa, and J. Dubuc. 2011. Reproductive tract defense and disease in postpartum dairy cows. Theriogenology 76:1610–1618.

Lee, A.R. 2018. An Evaluation of Physiological and Behavioral Indicators of Postpartum Diseases and Heat Stress in Dairy Cows.

Leutert, C., X. Von Krueger, J. Plöntzke, and W. Heuwieser. 2012. Evaluation of vaginoscopy for the diagnosis of clinical endometritis in dairy cows. J. Dairy Sci. 95:206–212.

Liboreiro, D.N., K.S. Machado, P.R.B. Silva, M.M. Maturana, T.K. Nishimura, A.P. Brandão, M.I. Endres, and R.C. Chebel. 2015. Characterization of peripartum rumination and activity of cows diagnosed with metabolic and uterine diseases. J. Dairy Sci. 98:6812–6827. doi:10.3168/jds.2014-8947.

Luo, G. 2016. A review of automatic selection methods for machine learning algorithms and hyper-parameter values. Netw. Model. Anal. Heal. Informatics Bioinforma. 5:18. doi:10.1007/s13721-016-0125-6.

Martiskainen, P., M. Järvinen, J.P. Skön, J. Tiirikainen, M. Kolehmainen, and J. Mononen. 2009. Cow behaviour pattern recognition using a three-dimensional accelerometer and support vector machines. Appl. Anim. Behav. Sci. 119:32–38. doi:10.1016/j.applanim.2009.03.005.

Matsui, K., and T. Okubo. 1991. A method for quantification of jaw movements suitable for use on free-ranging cattle. Appl. Anim. Behav. Sci. 32:107–116.

Mattachini, G., A. Antler, E. Riva, A. Arbel, and G. Provolo. 2013. Automated measurement of lying behavior for monitoring the comfort and welfare of lactating dairy cows. Livest. Sci. 158:145–150.

Mattachini, G., J. Pompe, A. Finzi, E. Tullo, E. Riva, and G. Provolo. 2019. Effects of feeding frequency on the lying behavior of dairy cows in a loose housing with automatic feeding and milking system. Animals 9. doi:10.3390/ani9040121.

Matthews, S.G., A.L. Miller, J. Clapp, T. Plötz, and I. Kyriazakis. 2016. Early detection of health and welfare compromises through automated detection of behavioural changes in pigs. Vet. J. 217:43–51. doi:10.1016/j.tvjl.2016.09.005.

Matthews, S.G., A.L. Miller, T. Plötz, and I. Kyriazakis. 2017. Automated tracking to measure behavioural changes in pigs for health and welfare monitoring. Sci. Rep. 7:1–12. doi:10.1038/s41598-017-17451-6.

McArt, J.A.A., D. V. Nydam, and G.R. Oetzel. 2012. Epidemiology of subclinical ketosis in early lactation dairy cattle. J. Dairy Sci. 95:5056–5066. doi:10.3168/jds.2012-5443.

McGowan, J.E., C.R. Burke, and J.G. Jago. 2007. Validation of a technology for objectively measuring behaviour in dairy cows and its application for oestrous detection. Page 136 in proceedings-New Zealand society of animal production. New Zealand Society of Animal Production; 1999.

Munksgaard, L., M.B. Jensen, L.J. Pedersen, S.W. Hansen, and L. Matthews. 2005. Quantifying behavioural priorities—Effects of time constraints on behaviour of dairy cows, Bos taurus. Appl. Anim. Behav. Sci. 92:3–14.

Neave, H.W., J. Lomb, D.M. Weary, S.J. LeBlanc, J.M. Huzzey, and M.A.G. von Keyserlingk. 2018. Behavioral changes before metritis diagnosis in dairy cows. J. Dairy Sci. 101:4388–4399. doi:10.3168/jds.2017-13078.

Opsomer, G., Y.T. Grohn, J. Hertl, M. Coryn, H. Deluyker, and A. de Kruif. 2000. Risk factors for postpartum ovarian dysfunction in high producing dairy cows in Belgium: a field study. Theriogenology 53:841–857.

Ouellet, V., E. Vasseur, W. Heuwieser, O. Burfeind, X. Maldague, and Charbonneau. 2016.

Evaluation of calving indicators measured by automated monitoring devices to predict the onset of calving in Holstein dairy cows. J. Dairy Sci. 99:1539–1548. doi:10.3168/jds.2015-10057.

Overton, M.W., W.M. Sischo, G.D. Temple, and D.A. Moore. 2002. Using time-lapse video photography to assess dairy cattle lying behavior in a free-stall barn. J. Dairy Sci. 85:2407–2413. doi:10.3168/jds.S0022-0302(02)74323-3.

Overton, T.R., and M.R. Waldron. 2004. Nutritional management of transition dairy cows: strategies to optimize metabolic health. J. Dairy Sci. 87:E105–E119.

Pahl, C., E. Hartung, A. Grothmann, K. Mahlkow-Nerge, and A. Haeussermann. 2014. Rumination activity of dairy cows in the 24 hours before and after calving. J. Dairy Sci. 97:6935–6941. doi:10.3168/jds.2014-8194.

Patbandha, T.K., T.K. Mohanty, S.S. Layek, A. Kumaresan, and K. Behera. 2012. Application of pre-partum feeding and social behaviour in predicting risk of developing metritis in crossbred cows. Appl. Anim. Behav. Sci. 139:10–17. doi:10.1016/j.applanim.2012.03.014.

Paudyal, S., F. Maunsell, J. Richeson, C. Risco, A. Donovan, and P. Pinedo. 2016. Peripartal rumination dynamics and health status in cows calving in hot and cool seasons. J. Dairy Sci. 99:9057–9068.

Paudyal, S., F.P. Maunsell, J.T. Richeson, C.A. Risco, D.A. Donovan, and P.J. Pinedo. 2018. Rumination time and monitoring of health disorders during early lactation. Animal 12:1484–1492. doi:10.1017/S1751731117002932.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M.

Perrot, and É. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. 12:2825–2830.

Pérez-Báez, J., T.V. Silva, C.A. Risco, R.C. Chebel, F. Cunha, A. De Vries, J.E.P. Santos, F.S. Lima, P. Pinedo, G.M. Schuenemann, R.C. Bicalho, R.O. Gilbert, S. Rodrigez-Zas, C.M. Seabury, G. Rosa, W.W. Thatcher, and K.N. K.N. Galvão. 2021. The economic cost of metritis in dairy herds. J. Dairy Sci. 104:3158–3168.

Piñeiro, J.M., B.T. Menichetti, A.A. Barragan, A.E. Relling, W.P. Weiss, S. Bas, and G.M. Schuenemann. 2019. Associations of pre- and postpartum lying time with metabolic, inflammation, and health status of lactating dairy cows. J. Dairy Sci. 102:3348–3361. doi:10.3168/jds.2018-15386.

Pleticha, S., M. Drillich, and W. Heuwieser. 2009. Evaluation of the Metricheck device and the gloved hand for the diagnosis of clinical endometritis in dairy cows. J. Dairy Sci. 92:5429–5435.

Potter, T.J., J. Guitian, J. Fishwick, P.J. Gordon, and I.M. Sheldon. 2010. Risk factors for clinical endometritis in postpartum dairy cattle. Theriogenology 74:127–134.

Probo, M., O.B. Pascottini, S. LeBlanc, G. Opsomer, and M. Hostens. 2018. Association between metabolic diseases and the culling risk of high-yielding dairy cows in a transition management facility using survival and decision tree analysis. J. Dairy Sci. 101:9419–9429. doi:10.3168/jds.2018-14422.

Proudfoot, K.L., J.M. Huzzey, and M.A.G. Von Keyserlingk. 2009. The effect of dystocia on the dry matter intake and behavior of Holstein cows. J. Dairy Sci. 92:4937–4944.

Proudfoot, K.L., D.M. Weary, and M.A.G. Von Keyserlingk. 2010. Behavior during transition

differs for cows diagnosed with claw horn lesions in mid lactation. J. Dairy Sci. 93:3970–3978.

R Core Team. 2017. R: A language and environment for statistical computing.

Radostits, O.M., C.C. Gay, K.W. Hinchcliff, and P.D. Constable. 2006. Veterinary Medicine E-Book: A Textbook of the Diseases of Cattle, Horses, Sheep, Pigs and Goats. Elsevier Health Sciences.

Rashid, A. 2003. Global information and early warning system on food and agriculture: appropriate technology and institutional development challenges. J. Zschau and A. Küppers, ed. Springer, Berlin, Heidelberg.

Royster, E., and S. Wagner. 2015. Treatment of Mastitis in Cattle. Vet. Clin. North Am. - Food Anim. Pract. 31:17–46. doi:10.1016/j.cvfa.2014.11.010.

Ruckebusch, Y. 1972. The relevance of drowsiness in the circadian cycle of farm animals. Anim. Behav. 20:637–643. doi:https://doi.org/10.1016/S0003-3472(72)80136-2.

Russell, J.B., and J.L. Rychlik. 2001. Factors that alter rumen microbial ecology. Science (80-.). 292:1119–1122.

Rutten, C.J., A.G.J. Velthuis, W. Steeneveld, and H. Hogeveen. 2013. Invited review: Sensors to support health management on dairy farms. J. Dairy Sci. 96:1928–1952. doi:10.3168/jds.2012-6107.

Saint-Dizier, M., and S. Chastant-Maillard. 2012. Towards an Automated Detection of Oestrus in Dairy Cattle. Reprod. Domest. Anim. 47:1056–1061. doi:10.1111/j.1439-0531.2011.01971.x.

Saint-Dizier, M., and S. Chastant-Maillard. 2018. Potential of connected devices to optimize cattle reproduction. Theriogenology 112:53–62. doi:10.1016/j.theriogenology.2017.09.033.

Saito, T., and M. Rehmsmeier. 2015. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One 10:1–21. doi:10.1371/journal.pone.0118432.

Sannmann, I., O. Burfeind, V. Suthar, A. Bos, M. Bruins, and W. Heuwieser. 2013. Evaluation of odor from vaginal discharge of cows in the first 10 days after calving by olfactory cognition and an electronic device. J. Dairy Sci. 96:5773–5779.

Sannmann, I., and W. Heuwieser. 2015. Intraobserver, interobserver, and test-retest reliabilities of an assessment of vaginal discharge from cows with and without acute puerperal metritis. J. Dairy Sci. 98:5460–5466.

Schirmann, K., N. Chapinal, D.M. Weary, L. Vickers, and M.A.G. Von Keyserlingk. 2013. Short communication: Rumination and feeding behavior before and after calving in dairy cows. J. Dairy Sci. 96:7088–7092. doi:10.3168/jds.2013-7023.

Schirmann, K., M.A.G. von Keyserlingk, D.M. Weary, D.M. Veira, and W. Heuwieser. 2009. Validation of a system for monitoring rumination in dairy cows. J. Dairy Sci. 92:6052–6055.

Schlageter-Tello, A., E.A.M. Bokkers, P.W.G. Groot Koerkamp, T. Van Hertem, S. Viazzi, C.E.B. Romanini, I. Halachmi, C. Bahr, D. Berckmans, and K. Lokhorst. 2014. Effect of merging levels of locomotion scores for dairy cows on intra- and interrater reliability and agreement. J. Dairy Sci. 97:5533–5542. doi:10.3168/jds.2014-8129.

Sepúlveda-Varas, P., D.M. Weary, and M.A.G. von Keyserlingk. 2014. Lying behavior and postpartum health status in grazing dairy cows. J. Dairy Sci. 97:6334–6343. doi:10.3168/jds.2014-8357.

Shahriar, M.S., D. Smith, A. Rahman, M. Freeman, J. Hills, R. Rawnsley, D. Henry, and G. Bishop-

Hurley. 2016. Detecting heat events in dairy cows using accelerometers and unsupervised learning. Comput. Electron. Agric. 128:20–26. doi:10.1016/j.compag.2016.08.009.

Sheldon, I.M., and H. Dobson. 2004. Postpartum uterine health in cattle. Anim. Reprod. Sci. 82:295–306.

Sheldon, I.M., G.S. Lewis, S. LeBlanc, and R.O. Gilbert. 2006. Defining postpartum uterine disease in cattle. Theriogenology 65:1516–1530. doi:10.1016/j.theriogenology.2005.08.021.

Sheldon, I.M., E.J. Williams, A.N.A. Miller, D.M. Nash, and S. Herath. 2008. Uterine diseases in cattle after parturition. Vet. J. 176:115–121.

Siivonen, J., S. Taponen, M. Hovinen, M. Pastell, B.J. Lensink, S. Pyörälä, and L. Hänninen. 2011. Impact of acute clinical mastitis on cow behaviour. Appl. Anim. Behav. Sci. 132:101–106. doi:10.1016/j.applanim.2011.04.005.

Soriani, N., G. Panella, and L. Calamari. 2013. Rumination time during the summer season and its relationships with metabolic conditions and milk production. J. Dairy Sci. 96:5082–5094. doi:10.3168/jds.2013-6620.

Soriani, N., E. Trevisi, and L. Calamari. 2012. Relationships between rumination time, metabolic conditions, and health status in dairy cows during the transition period. J. Anim. Sci. 90:4544–4554.

Stangaferro, M.L., R. Wijma, L.S. Caixeta, M.A. Al-Abri, and J.O. Giordano. 2016a. Use of rumination and activity monitoring for the identification of dairy cows with health disorders: Part III. Metritis. J. Dairy Sci. 99:7422–7433. doi:10.3168/jds.2016-11352.

Stangaferro, M.L., R. Wijma, L.S. Caixeta, M.A. Al-Abri, and J.O. Giordano. 2016b. Use of rumination and activity monitoring for the identification of dairy cows with health disorders:

Part I. Metabolic and digestive disorders. J. Dairy Sci. 99:7395–7410. doi:10.3168/jds.2016-10907.

Stangaferro, M.L., R. Wijma, L.S. Caixeta, M.A. Al-Abri, and J.O. Giordano. 2016c. Use of rumination and activity monitoring for the identification of dairy cows with health disorders: Part I. Metabolic and digestive disorders. J. Dairy Sci. 99:7395–7410.

Steensels, M., A. Antler, C. Bahr, D. Berckmans, E. Maltz, and I. Halachmi. 2016. A decision-tree model to detect post-calving diseases based on rumination, activity, milk yield, BW and voluntary visits to the milking robot. Animal 10:1493–1500. doi:10.1017/S1751731116000744.

Steensels, M., E. Maltz, C. Bahr, D. Berckmans, A. Antler, and I. Halachmi. 2017. Towards practical application of sensors for monitoring animal health: The effect of post-calving health problems on rumination duration, activity and milk yield. J. Dairy Res. 84:132–138. doi:10.1017/S0022029917000176.

Stoye, S., M.A. Porter, and M. Stamp Dawkins. 2012. Synchronized lying in cattle in relation to time of day. Livest. Sci. 149:70–73. doi:10.1016/j.livsci.2012.06.028.

Sturm, V., D. Efrosinin, M. Öhlschuster, E. Gusterer, M. Drillich, and M. Iwersen. 2020. Combination of Sensor Data and Health Monitoring for Early Detection of Subclinical Ketosis in Dairy Cows. Sensors 20. doi:10.3390/s20051484.

Tamura, T., Y. Okubo, Y. Deguchi, S. Koshikawa, M. Takahashi, Y. Chida, and K. Okada. 2019. Dairy cattle behavior classifications based on decision tree learning using 3-axis neck-mounted accelerometers. Anim. Sci. J. 90:589–596. doi:10.1111/asj.13184.

Thompson, R.N., and E. Brooks-Pollock. 2019. Preface to theme issue “Modelling infectious

disease outbreaks in humans, animals and plants: Epidemic forecasting and control". Philos. Trans. R. Soc. B Biol. Sci. 374. doi:10.1098/rstb.2019.0375.

Tizard, I. 2008. Sickness behavior, its mechanisms and significance.. Anim. Health Res. Rev. 9:87–99. doi:10.1017/S1466252308001448.

Tremblay, M., M. Kammer, H. Lange, S. Plattner, C. Baumgartner, J.A. Stegeman, J. Duda, R. Mansfeld, and D. Döpfer. 2018. Identifying poor metabolic adaptation during early lactation in dairy cows using cluster analysis. J. Dairy Sci. 101:7311–7321. doi:https://doi.org/10.3168/jds.2017-13582.

Tsai, I.C. 2017. Differences in Behavioral and Physiological Variables Measured With Precision Dairy Monitoring Technologies Associated With Postpartum Diseases.

Urton, G., M.A.G. Von Keyserlingk, and D.M. Weary. 2005. Feeding behavior identifies dairy cows at risk for metritis. J. Dairy Sci. 88:2843–2849. doi:10.3168/jds.S0022-0302(05)72965-9.

Vanrell, S.R., J.O. Chelotti, J.R. Galli, H.L. Rufiner, and D.H. Milone. 2014. 3D Acceleration for Heat Detection in Dairy Cows. Sexto Congr. Argentino Agroinformatica 64–73.

Vapnick, V. 1995. The Nature of Statistical Learning Theory. Springer, New York.

Vasseur, E., J. Rushen, D.B. Haley, and A.M. de Passillé. 2012. Sampling cows to assess lying time for on-farm animal welfare assessment. J. Dairy Sci. 95:4968–4977. doi:10.3168/jds.2011-5176.

Vidal, G., J. Sharpnack, P. Pinedo, I.C. Tsai, and A.R. Lee. Comparative performance analysis of three machine learning algorithms applied to sensor data in dairy cattle to predict metritis events. doi:doi:10.7910/DVN/BZX8KD.

Wagner, N., V. Antoine, M.-M. Mialon, R. Lardy, M. Silberberg, J. Koko, and I. Veissier. 2020.

Machine learning to detect behavioural anomalies in dairy cows under subacute ruminal acidosis. Comput. Electron. Agric. 170:105233. doi:https://doi.org/10.1016/j.compag.2020.105233.

Walsh, R.B., J.S. Walton, D.F. Kelton, S.J. LeBlanc, K.E. Leslie, and T.F. Duffield. 2007. The effect of subclinical ketosis in early lactation on reproductive performance of postpartum dairy cows. J. Dairy Sci. 90:2788–2796. doi:10.3168/jds.2006-560.

Walton, E., C. Casey, J. Mitsch, J.A. Vázquez-Diosdado, J. Yan, T. Dottorini, K.A. Ellis, A. Winterlich, and J. Kaler. 2018. Evaluation of sampling frequency, window size and sensor position for classification of sheep behaviour. R. Soc. Open Sci. 5. doi:10.1098/rsos.171442.

Wang, Q., Y. Ma, K. Zhao, and Y. Tian. 2020. A Comprehensive Survey of Loss Functions in Machine Learning. Ann. Data Sci.. doi:10.1007/s40745-020-00253-5.

Wathes, C.M., H.H. Kristensen, J.-M. Aerts, and D. Berckmans. 2008. Is precision livestock farming an engineer's daydream or nightmare, an animal's friend or foe, and a farmer's panacea or pitfall?. Comput. Electron. Agric. 64:2–10.

Weary, D.M., J.M. Huzzey, and M.A.G. Von Keyserlingk. 2009. Board-invited Review: Using behavior to predict and identify ill health in animals. J. Anim. Sci. 87:770–777. doi:10.2527/jas.2008-1297.

Wehrend, A., K. Failing, and H. Bostedt. 2003. Cervimetry and ultrasonographic observations of the cervix regression in dairy cows during the first 10 days post partum. J. Vet. Med. Ser. A 50:470–473.

White, R.R., M.B. Hall, J.L. Firkins, and P.J. Kononoff. 2017. Physically adjusted neutral detergent fiber system for lactating dairy cow rations. I: Deriving equations that identify factors that

influence effectiveness of fiber. J. Dairy Sci. 100:9551–9568. doi:10.3168/jds.2017-12765.

Wickham, H. 2009. Ggplot2: Elegant Graphics for Data Analysis. Springer, New York.

Williams, M.L., N. Mac Parthaláin, P. Brewer, W.P.J. James, and M.T. Rose. 2016. A novel behavioral model of the pasture-based dairy cow from GPS data using data mining and machine learning techniques. J. Dairy Sci. 99:2063–2075. doi:10.3168/jds.2015-10254.

Wolfger, B., E. Timsit, E.A. Pajor, N. Cook, H.W. Barkema, and K. Orsel. 2015. Technical note: Accuracy of an ear tag-attached accelerometer to monitor rumination and feeding behavior in feedlot cattle. J. Anim. Sci. 93:3164–3168. doi:10.2527/jas.2014-8802.

Wolpert, D.H. 1996. The lack of a priori distinctions between learning algorithms. Neural Comput. 8:1341–1390.

Zebeli, Q., M. Tafaj, H. Steingass, B. Metzler, and W. Drochner. 2006. Effects of physically effective fiber on digestive processes and milk fat content in early lactating dairy cows fed total mixed rations. J. Dairy Sci. 89:651–668. doi:10.3168/jds.S0022-0302(06)72129-4.